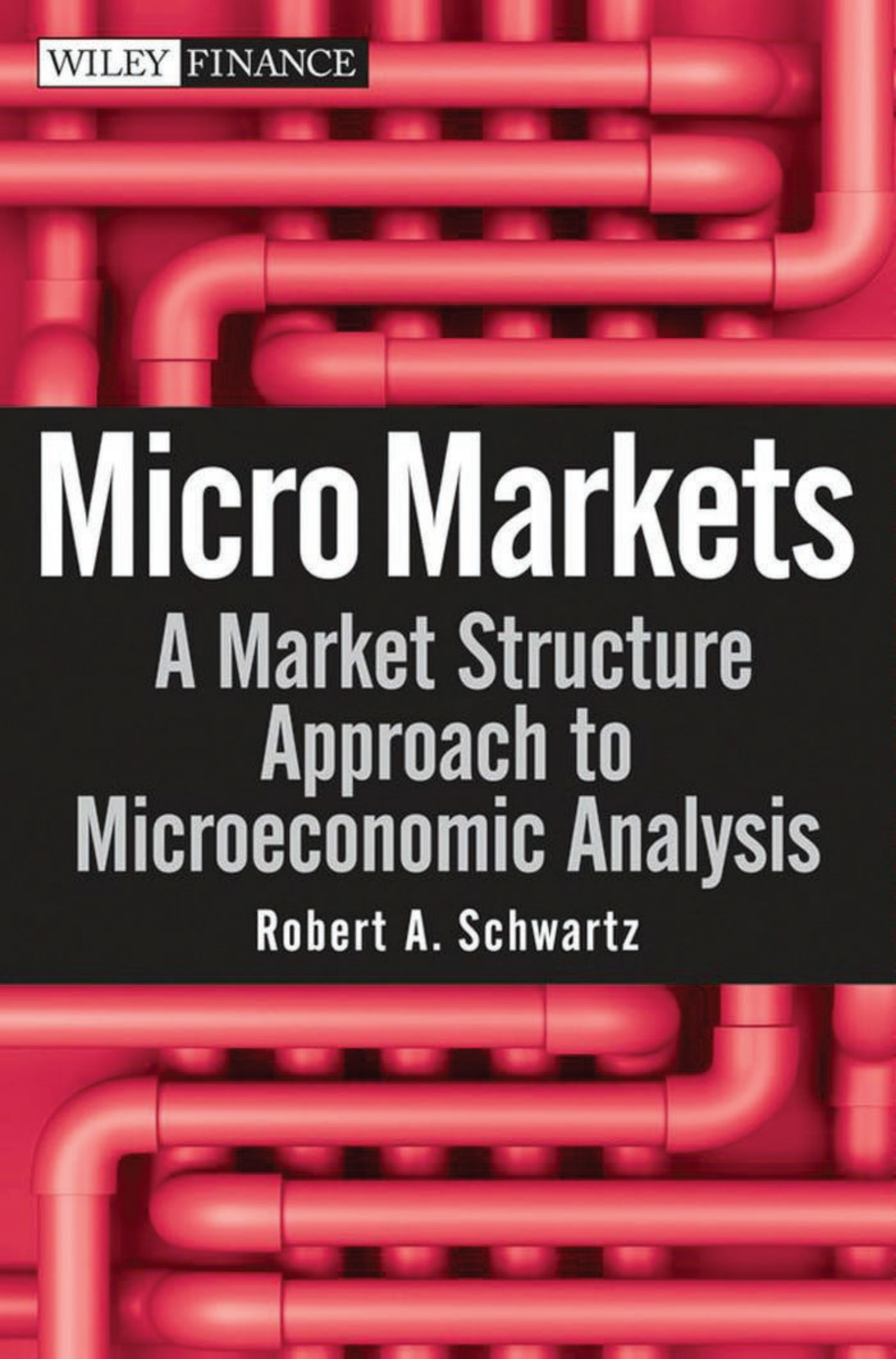


WILEY FINANCE

# Micro Markets

## A Market Structure Approach to Microeconomic Analysis

Robert A. Schwartz

# Micro Markets

Founded in 1807, John Wiley & Sons is the oldest independent publishing company in the United States. With offices in North America, Europe, Australia and Asia, Wiley is globally committed to developing and marketing print and electronic products and services for our customers' professional and personal knowledge and understanding.

The Wiley Finance series contains books written specifically for finance and investment professionals as well as sophisticated individual investors and their financial advisors. Book topics range from portfolio management to e-commerce, risk management, financial engineering, valuation and financial instrument analysis, as well as much more.

For a list of available titles, please visit our Web site at www.WileyFinance.com.

# Micro Markets

## *A Market Structure Approach to Microeconomic Analysis*

ROBERT A. SCHWARTZ

WITH MICHAEL G. CAREW AND

TATIANA MAKSIMENKO

WILEY

John Wiley & Sons, Inc.

Copyright © 2010 by Robert A. Schwartz. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.wiley.com.

***Library of Congress Cataloging-in-Publication Data:***

Schwartz, Robert A. (Robert Alan), 1937-
Micro markets : a market structure approach to microeconomic analysis /
Robert A. Schwartz.
p. cm. – (Wiley finance series)
Includes index.
ISBN 978-0-470-44765-9 (cloth)
1. Stock exchanges. 2. Microeconomics. I. Title.
HG4551.S3934 2010
338.5–dc22 2009042659

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

*To my family, Jody, Emily, and Leigh*

# Contents

**Preface** **xi**

Overview of the Book xiv
Related Publications xvii

**Acknowledgments** **xix**

**CHAPTER 1**
**Introduction to Market-Driven Economics** **1**

Theory and Reality 2
Primary and Secondary Markets 3
Frictionless versus Nonfrictionless Markets 5
How Well Do Markets Function? 6
A Market Is an Ecology 10
Basic Microeconomic Concepts 11
Frictionless Markets, the Gulf Stream, and a Day at the Races 20
Chapter Summary 21
Questions to Think About 23
TraderEx Simulation: Let's Look at a Market 24

**CHAPTER 2**
**The Consumer Choice Model: What Would You Really Like to Do?** **27**

The Consumer Choice Model and the *X*, *Y* Trade-Off 28
The Consumer Choice Model and the Current Consumption, Future Consumption Trade-Off 35
The Consumer Choice Model and the Risk-Return Trade-Off 43
Comparative Statics 64
Chapter Summary 73
Let's Consider an Application: Philanthropy 76

TraderEx Simulation: What Are Your Attitudes Toward Risk? 78

## CHAPTER 3
## Demand Meets Supply 81

Should Have, Would Have, and Did: The Decisions of a Representative Consumer 82
The Demand Curve for *X* 84
The Sensitivity of the Demand for *X* to the Price of *X*, the Price of *Y*, and Income 86
From Individual Demand Curves to Market Demand Curves 96
The Demand to Hold Shares of a Risky Asset 99
The Investor's Demand Curve for the Risky Asset 100
Equilibrium in the Market for *X* 102
Equilibrium in the Market for Shares of the Risky Asset 109
Market Equilibrium 112
Stocks, Cakes, and Substitutability 116
Chapter Summary 122
Let's Consider an Application: Bandwagon and Snob Effects 124
TraderEx Simulation: Call Market Trading 127

## CHAPTER 4
## Microeconomic Analysis Goes to Market 131

Short-Run Demand and Supply for Product *X* 133
Adjustment Costs 136
Liquidity 142
Market Structure 148
The Gains from Trading 164
Strategic Order Placement 170
The Big Picture 176
Chapter Summary 178
Let's Consider an Application: Why Bid-Ask Spreads Exist in Securities Markets 180
TraderEx Simulation: What Are Your Attitudes Toward Trading Costs? 184

## CHAPTER 5
## Supply and the Costs of Production 187

The Production Function for *X* 188

Cost Curves 193
Determining the Best Profit Output 206
Supply 208
Application to a Dealer Firm 215
From General Principles to the Details of a Specific Market 230
Chapter Summary 231
Let's Consider an Application: The Optimal Schooling Behavior of Fish 234
TraderEx Simulation: What Is It Like to Be a Securities Dealer? 237

**CHAPTER 6**
**Sources and Nature of Competition** **241**

Perfect Competition 242
Monopoly 249
Imperfect Competition: Monopolistic Competition 260
Imperfect Competition: Oligopoly 263
How Competition Plays Out in the Equity Markets 270
Importance of Price as a Competitive Variable 281
Chapter Summary 282
Let's Consider an Application: Why Firms Extend Trade Credit 287
TraderEx Simulation: Intermarket Competition in the Exchange Space 289

**CHAPTER 7**
**Market Efficiency** **293**

Allocation Efficiency 295
Market Failure 298
Information and Expectations 305
Info about Information 307
Four Dimensions of Informational Efficiency 311
A Test of Market Efficiency 322
Price Discovery Efficiency 325
The Big Picture on Financial Market Efficiency 330
Markets at Risk: The Financial Turmoil of 2007–2009 333
The Greater Generality of the Frictions Problem 341
Chapter Summary 343
Let's Consider an Application: How Are Equilibrium Prices Discovered in a Competitive Market? 347
TraderEx Simulation: Finding an Equilibrium Value 351

**CHAPTER 8**
**Public Policy and the Interplay between Competition, Technology, and Regulation** **353**

Governance Structure and Self-Regulation 355
Technology, Automation, and the NASDAQ Stock Market 360
The Evolving Regulatory Focus 368
Governmental Regulatory Intervention 373
Regulation of the U.S. Equity Markets 375
Caveats for Public Policy 397
Chapter Summary 401
Questions to Think About 406
TraderEx Simulation: Economic Effects of an Order Protection Rule 410

**Glossary** **413**

**About the Author** **429**

**Index** **431**

# Preface

What is a micro market, and what is microeconomics all about? In a nutshell, a *micro market* is the market for a specific good, service, factor of production, or asset (in contrast to the *macro market*, which is all of a country's micro markets in aggregate). *Microeconomics* is about how households, firms, and industries interact in the marketplace to determine the production, pricing, and distribution of society's scarce resources.

This book is about how micro markets work. It is about the forces that drive production and consumption decisions. It is about the level of economic efficiency that markets and the people operating in them can achieve. While our focus is on micro markets, our subject is microeconomic theory. Microeconomic theory is a powerful and compelling body of thought that provides much insight into how the micro markets operate.

Here are some of the basic microeconomic issues that we will be addressing.

- **Consumption decisions:** How are these decisions made? That is, how can a household best allocate its scarce resources (its income and wealth) to maximize its well-being (or, as we like to call it, *its utility*).
- **Production decisions:** How are these decisions made? That is, how can a firm best use its factors of production, establish its output rate, and, if it has the power to do so, set a price for its product that will maximize its well-being (or, as we like to call it, *its profits*)?
- **Industry output:** How do the decisions of firms aggregate to the output of the industries that the firms are in?
- **Firm and industry equilibrium:** How does equilibria for firms and their industries depend on the competitive structure of the industry (meaning, is the industry very competitive, is it a monopoly, or is it something else)?

You will appreciate the tremendous importance that prices play in microeconomic analysis. Prices are the signals, the traffic lights, the force that the moral philosopher Adam Smith called "an invisible hand" that directs the allocation of resources in a free and competitive market. This invisible hand offers carrots: profits for successful firms and superior consumption choices for clever households. It also carries a stick, and it is a harsh taskmaster: inefficient firms fail, inefficient employees are fired, from time to time the macro economy will hit a speed bump and spiral down, and innocent firms and people suffer along with the guilty.

The U.S. economy survived a tech (dot-com) bubble at the start of the century, but by 2007 we were facing something far more serious: a housing bubble, a plethora of new and complex financial instruments that relatively few people really understood, and leverage rates that had reached sky-high levels. The storm clouds had started to gather in the summer of 2007. In early August, the first signs of a severe credit crunch appeared, as several major issuers of asset-backed commercial paper, which had provided prior funding of mortgage-backed securities, found that they were no longer able to refund their outstanding issues. This evaporation of liquidity spread to other sectors of the capital markets. Over the ensuing year, the illiquidity problem aggravated and, as it did, some major financial institutions, such as Bear Stearns, weakened. In the fall of 2008, a major financial crisis erupted, and it quickly spread to the broader markets in the United States and beyond.

Here are some highlights: Lehman Brothers, the venerable investment banking firm, totally collapsed; the big insurance company, American International Group (AIG), along with the two giant mortgage companies, Fannie Mae and Freddie Mac, went on government life support; Merrill Lynch is now part of Bank of America, and Wachovia has been acquired by Wells Fargo. Goldman Sachs and Morgan Stanley are being regulated by the Federal Reserve as bank holding companies, while Citicorp and Bank of America increasingly have come under government control.

These developments profoundly impacted the broader economy. As credit lines froze and more nonfinancial firms headed toward bankruptcy, it became crystal clear that the financial sector is not simply a conduit for transmitting funds from those who save to those who invest. Rather, the finance function is complex, it is interconnected, and it pervades the micro markets. As one frightening development led to another and 2008 turned into 2009, systemic risk came to be widely feared.

What does this say about Adam Smith's invisible hand, that unseeable force that guides people to do what is best for others while striving to do what is best for themselves? Technology changes, competition, and appropriate regulatory interventions can move a macro economy forward. Nevertheless, it is important also to recognize that change in a dynamically evolving environment forces many people (along with firms and government) to make big adjustments. How well do they do it? Are they adequately educated, sufficiently trained, mentally prepared, and financially capable of making the adjustments? Perhaps in the long run Adam Smith's invisible hand moves society in a desirable direction but, as we live our lives in the short run, additional factors come into play. We must take a closer look at issues such as:

- **Market frictions:** When firms implement their production decisions and households implement their consumption decisions in real-world markets, choices that are made and the outcomes that are realized are affected by transaction costs, blockages, risks, uncertainties, imperfect knowledge, and other impediments. We call these impediments *frictions*. How do the frictions affect market outcomes?
- **Efficiency of a free market:** We will consider the ability of the micro markets to efficiently serve our needs in a free-market setting. Might the free market fail to deliver the results that we want from a public policy perspective? Are there causes of market failure and, if so, what are they?
- **Efficiency of government intervention:** If a market fails to perform with reasonable efficiency, what should a government regulatory authority do about it? How efficient is government intervention itself? What is the appropriate balance between free-market competition and government intervention? Do not expect a definitive answer to this one, but be assured that we will pay attention to this age-old public policy question.

Using microeconomic analysis to come to grips with these issues requires both science and art. It involves science because of the rigor that we demand of our theories and the precision of the econometric methods that we use to test our theories and to assess our predictions. It involves art because, all said and done, theories are based on simplifying assumptions and thus do not conform exactly to reality. Moreover, we run our econometric tests on data that, to at least some extent, are incomplete and imperfect. Thus judgment is called for and, in applying it, economists often disagree with one another (ask any three economists a question and you might get four opinions).

When personal judgment comes into play, personal biases enter the picture. That is fine, it is reality. Personal opinions should be put on the table and discussed. But it is best to be open about one's biases. So, here is one of mine, and it is no doubt reflected in this book. Appropriate regulation is of course needed, but like many microeconomists I generally favor a free market stance. At times I have been rather critical of some of the government regulations that have been applied to the architectural structure of the U.S. securities markets (one of my primary fields of research). Why is this? While I believe that from a microeconomics perspective the securities markets are in some important respects quite imperfect, I also observe that government regulatory intervention can have its own undesired, unexpected, and unintended consequences. For this reason, I retain my faith in the comparative ability of freely competitive markets to produce desirable results in the long run.

There are of course no absolute answers for all of this. I hope that your study of microeconomics will help to sharpen your own thinking about these important issues that economics as a science has long been striving to come to grips with.

## OVERVIEW OF THE BOOK

Microeconomic theory provides an excellent set of tools for addressing our list of issues and questions. As we present these tools, we set forth our analysis with regard to broadly defined, abstract resources (the representative household, for instance, consumes just two generic products, $X$ and $Y$). We then apply a number of these tools to a specific market: a market where the already-issued shares of corporate stock trade (that is, an equity market).

The equity markets are vitally important to the broad economy, they are intriguing to study, they provide a unified set of applications that we use throughout the book, and microeconomic analysis offers much insight into their operations. As we proceed, we intersperse relatively abstract discussions of the fundamental microeconomic formulations with more concrete analyses of the equity markets.

The book comprises eight chapters:

**Chapter 1: Introduction to Market-Driven Economics.** In this chapter we present basic thoughts (theory versus reality, frictionless versus nonfrictionless markets, and the idea that a market can be viewed as an ecology) and concepts [*ceteris paribus* (all things being equal), fixed costs versus variable costs, long-run versus short-run analysis, equilibrium, marginal analysis, elasticity, and the difference between maximum, minimum, and optimum] that are needed for our subsequent analyses.

**Chapter 2: The Consumer Choice Model: What Would You Really Like to Do?** This chapter establishes the theoretical framework for obtaining a household's demand for consumption goods and services, and its demand to hold shares of a risky financial asset. We present the analysis for our two generic goods ($X$ and $Y$), for two more defined consumption bundles (present and future consumption), and for two attributes of a financial asset (risk and return).

**Chapter 3: Demand Meets Supply.** This chapter builds on the framework established in Chapter 2 to obtain the individual household's demand curve and the aggregate market (industry) demand curve for the generic product X, and for shares of the risky financial asset. The chapter also shows how we use a measure of responsiveness called *elasticity* to represent the sensitivity of one variable (such as the quantity of *X* demanded) to another variable (such as the price of *X*). We introduce a supply curve and show how the intersection of demand and supply in a competitive market establishes equilibrium values for our two key variables (price and quantity).

**Chapter 4: Microeconomic Analysis Goes to Market.** We enter the nonfrictionless environment in this chapter. The following key concepts are put forth: short-run versus long-run analysis, the effect of adjustment costs on market outcomes, the liquidity (or lack thereof) of a marketplace, the architectural structure of a marketplace, the gains participants realize from trading, and their strategic order placement decisions.

**Chapter 5: Supply and the Costs of Production.** In this chapter we change our focus from households to firms and, in so doing, derive the supply curve for a competitive firm and industry. To do this, we delve into the production function for the generic product *X* and the cost curves for producing *X*, show how a firm determines the most profitable amount of *X* to produce and, knowing all of this, obtain the supply curve of *X* for the competitive firm. We then apply the analysis to a specific kind of firm: an equity dealer firm.

**Chapter 6: Sources and Nature of Competition.** Building further on the analysis developed in Chapter 5, we show in this chapter how the output and pricing decisions of firms and aggregate industry output depend on the competitive structure of the industry that the firms are operating in (highly competitive, a monopoly, or something else). We then take a look at how competition plays out in the equity markets. The chapter concludes with a consideration of the importance of price as a competitive variable in a fiercely competitive industry (other variables such as advertising and product quality also have major roles to play).

**Chapter 7: Market Efficiency.** By now, key building blocks have been put on the table. In this chapter we pull them together and ask a bottom-line question—how capable is the free market in getting us the outcomes that, as a body politic, we want? In other words, how efficient is a free, competitive market? We turn our attention to various causes of market failure, giving particular emphasis to the enormous complexity, imprecision, and incompleteness of much of the information that we base our production, consumption, and investment decisions on. We consider the financial market turmoil that started in 2007 and roared through the economy in 2008 and 2009, and we will possibly leave you pondering (perhaps with a favorable eye) the desirability of bringing additional government regulation into the picture.

**Chapter 8: Public Policy and the Interplay between Competition, Technology, and Regulation.** In this concluding chapter we turn primarily to issues concerning government involvement in market outcomes. We first pay attention to a firm's governance structure (namely, whether it is a membership organization or a for profit operation) and to the self-regulatory obligations of a firm such as a securities exchange. We consider the impact of technology on the economy, underscore how technology change seriously challenges the regulators (by giving rise to new regulatory issues that are difficult to understand and not easy to keep up with), and turn to government's role in regulating the micro markets. We summarize the evolving regulatory focus with reference to three items: honesty, the efficiency of the market for broker/dealer services, and a market's efficiency with regard to the pricing of equity shares. The chapter sets forth various reasons why a government's regulatory intervention can itself be flawed (this discussion parallels our focus in Chapter 7 on why the free market may itself fail to deliver fully efficient results). A substantial portion of the chapter is then devoted to six specific equity market initiatives that have been taken by the U.S. regulatory authorities since 1975.

Our discussion in Chapter 8 emphasizes that public policy issues can be enormously complex, and that government intervention is far from an easy and complete panacea. For those who favor government intervention in the workings of the micro markets, we offer a number of caveats. We then end the book with this thought: While the free market may be far from perfect, we should not lose faith in its comparative efficacy to ultimately deliver socially desirable results.

Three items are included at the end of each chapter: a chapter summary, an application that includes questions for you to think about, and a computer exercise that you can do if you are in possession of *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course* (John Wiley & Sons, 2010). These items are purposed to further your comprehension of the material and to give you additional experience applying the thoughts and analyses that we present in this book. Further, with *Mastering the Art of Equity Trading*, you will be able to access the TraderEx simulation software via a Wiley web site.

A glossary of key terms is included at the end of the book.

## RELATED PUBLICATIONS

*Micro Markets Workbook*, by Robert A. Schwartz, Michael G. Carew, and Tatiana Maksimenko provides a means for reviewing the materials presented in the text and for enhancing your understanding of the principles of microeconomic theory.

Our text, as we have pointed out, devotes considerable attention to the equity markets. Hands-on experience with trading in an equity market can be obtained with the use of *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course,* by Robert A. Schwartz, Gregory Sipress, and Bruce Weber (John Wiley & Sons, 2010). TraderEx is a computerized model of an equity market that enables a user to trade in a simulated environment, and the TraderEx book and the software that comes with it allows you to actually experience market operations. As noted in the previous section, the end of chapter material in each of the eight chapters of our *Micro Markets* text includes a computer simulation exercise that can be undertaken with the software and additional instructions that are provided in the TraderEx book.

Some of the material in this book draws from Robert A. Schwartz and Reto Francioni, *Equity Markets in Action: The Fundamentals of Liquidity, Market Structure, and Trading* (John Wiley & Sons, 2004) and from Robert A. Schwartz, Reto Francioni, and Bruce Weber, *The Equity Trader Course* (John Wiley & Sons, 2006). Both of these books provide further information about equity market structure, the role and operations of exchanges, and a spectrum of other topics including trading costs, trading strategies, public policy, and government regulatory intervention.

# Acknowledgments

My first and greatest thanks are extended to Michael Carew, my colleague at Baruch, and to Tatiana Maksimenko a doctoral student at Baruch, who have in many ways participated in the production of this book, and have provided me with continual support and encouragement during the production process. I am also deeply grateful to Antoinette Colaninno, my assistant at Baruch, for all the help that she has given me. My friend and coauthor, Jacob Paroush, and my colleague, Norman Klienberg, have carefully reviewed the manuscript and have offered very helpful feedback. One other Baruch doctoral student, Sila Saylak, has spent long hours carefully reading the manuscript, giving me excellent feedback and advice, and helping with the production of the exhibits. Pamela van Giessen and Emilie Herman, my editors at John Wiley & Sons, have been a delight to work with, and I am most appreciative of the excellent advice and support that they have consistently given me. I also extend my heartfelt thanks to a number of other people, including two excellent research assistants, Bhavesh Dalal and Deepak Mulani, and participants in my doctoral seminar at Baruch, Mert Demir, Yalan Feng, Huajing Hu, Jeff Hu, Jun Jiang, Sibel Korkmaz, Jin Lei, Yili Lian, Scott Murray, Ko Woon Oh, Jangwon Suh, Ti-Jen Tsao, and Sha Zhao. One other person must be mentioned: I am unendingly thankful for the knowledge provided to me many years ago by my professor, dissertation advisor, and mentor, Gary Becker.

# CHAPTER 1

# Introduction to Market-Driven Economics

Markets are all around us. They lie at the center of our economic lives. They tie people together, they tie countries together, and, today more than ever, they are tying the world together in one global economy. Our subject, microeconomics, is about the functioning of a market. Along with being of critical importance to all of us, markets are exciting to study.

The book's title is *Micro Markets*. By "micro" we are referring to the individual markets that comprise a broad economy. In contrast, the term "macro" refers to markets in the aggregate (that is, the broad economy). A micro market may be the market for a specific product (such as the book that you are now holding), or for a specific service (education, for example), or for a financial asset (e.g., a stock or bond). Have you ever considered how a micro market might operate? Have you ever thought about the forces that drive the micro markets, about how the markets can interact with each other, about who the key players in a marketplace might be, about how output might be set, consumption or asset holdings determined, and prices established? I bet you have. The objective of this book is to help formalize your insights into the dynamic behavior of the micro markets.

It is easy to take many of our markets for granted. When going to the supermarket with a shopping list, most of us simply buy our butter, eggs, and beer. But when the price of gas at the pump more than triples or falls by over thirty percent, we start asking questions. We ask questions when a bubble in the housing market bursts. And many of us scratch our heads when the Dow Jones Industrial average drops over 300 points or soars over 300 points in a single day. The fact is, markets are complex, the forces that impact them are intricate, and market breakdowns can occur. Just how efficiently a market operates depends on how it is structured. Very importantly, alternative structures exist. A marketplace can be designed, and the people who do this can be thought of as *market architects*. The market architects can get guidance from microeconomic theory.

In this introductory chapter we consider the relationship between theory and reality, between frictionless and nonfrictionless markets, and suggest that a market can be viewed as an ecology. We identify important economic vocabulary that is needed for our subsequent analyses: *ceteris paribus* (a Greek term meaning "all else constant" that we explain shortly), fixed versus variable costs and returns, long-run versus short-run analysis, the concept of an equilibrium, marginal analysis, elasticity (a measure of responsiveness that we explain shortly), and the difference between "maximum," "minimum," and "optimum."

## THEORY AND REALITY

Microeconomic *theory* is just that, it is theory. As such, it may appear to be a highly abstract, overly simplified, unpalatably unrealistic representation of any actual market. Imagine if a Martian's first sighting of a human being was a Picasso portrait. Quite startling, wouldn't you say? So why was Picasso a great painter? Here is one reason: his abstractions were great. He was able to capture the essence of a person without delivering an exact replication of the subject. That too is the goal of microeconomic theory: via the use of simplifying (and albeit often unrealistic assumptions), to capture the essence of an issue.

Real-world markets can be extremely intricate. Their many subtleties and extensive institutional detail make them almost impossible to understand in all their glory. Those who have worked in an industry can start to understand it, but even then their comprehension is typically intuitive. Ask somebody who has had the requisite experience in an industry to describe it and you will generally get just that—a great deal of description. What can we learn from the description? How can we classify thinking, generalize, and apply the knowledge more broadly to other industries and markets? That is where theory has a role to play.

The process of theorizing involves making the simplifying abstractions so as to get to the heart of the matter without being overwhelmed with unnecessary detail. The hallmark of a good theory is that, through the intelligent use of assumptions and analysis, it has honed in on the question that we are interested in. In other words, a good theory has sliced away inconsequential detail so as to gain the insights that we want.

You might appreciate microeconomic theory with real interest only when you see it being applied to an actual market. So let's do that. Throughout much of this book, we present microeconomic theory with reference to one specific, real-world micro market: the equity market. This is an excellent market to analyze. Equities are a critically important financial asset for scores of investors, corporations, and the macro economy. Corporate equities (also referred to as ‘‘stocks’’) represent shares of ownership in public companies and, as such, equity financing is a major source of the financial capital that firms must have to undertake their operations. On a national level, equities comprise a major part of the portfolios of both individual investors and institutional investors such as mutual funds and pension funds. Well-functioning equity markets are essential to the vibrancy and growth of a macro economy. And in light of its dynamic properties, an equity market is a particularly intriguing micro market to study.

While the equity markets as part of the overall capital markets are our exemplative focus, other markets for consumer goods and *factors of production* have their own critically important and distinctive characteristics. Consumer items include services, and goods that can be classified as either durable or nondurable. The classic factors of production are land, labor, and physical capital. Human capital has also been treated as a factor of production in its own right. In a nonfrictionless environment we might additionally include financial capital as a productive factor (in a frictionless world, physical inputs can be turned into physical outputs without the use of any financial instrument). Further, if the nonfrictionless environment is characterized by imperfect (incomplete) information, information itself is a productive factor. All of these markets can have their own particular architecture and rules.

## PRIMARY AND SECONDARY MARKETS

Comprehending a subject such as microeconomics requires understanding new vocabulary and, throughout much of this book, we will be introducing new terms. In essence, you will be learning a new language. Language lies at the heart of thought—without it, our ability to think and to communicate would be severely impaired. We suggest that you reflect on the terms that we introduce. From time to time, try using some of them in daily conversation with your friends. We do, but then again our noneconomist friends often do not understand what we are saying.

Here are two terms that apply to the equity markets: the *primary market* and the *secondary market*. The primary market for equity shares is the market where new shares are issued. The secondary market is where already issued shares are traded. A corporation looking to raise financial capital will issue new shares with the assistance of a firm like Goldman Sachs or Morgan Stanley. If the firm is going public and is issuing shares for the first time, the issuance is called an initial public offering (IPO). Will the IPO be successful? Will individual retail investors, pension funds, mutual funds, and other institutional investors actually buy the shares? The answer depends on a number of factors, including the price the shares are offered at, the risk an investment in them would involve, and the company's future prospects for earnings growth.

One other consideration is paramount: If you buy the shares when they are offered in the primary market, will you be able to sell them whenever you want in a secondary market? Perhaps you will need funds to meet an expense, or will wish to switch into some other investment. Will you be able to raise the cash by selling the shares? Will your share holdings be adequately *liquid*?

"Liquidity" is something that we all know something about. Here is a quick and easy definition: an asset is liquid if it can be bought or sold reasonably quickly, at a reasonable price, in reasonably large amounts, in a reasonably short period of time. Please do not ask what "reasonable" means—if you do, we might have to answer, "It depends." Regardless, by this definition, it is clear that the equity shares of most large, publicly traded corporations are relatively liquid assets and, in comparison, that a house or a condominium apartment is a very illiquid asset.

Existing share holdings (and homes) can be bought and sold in a secondary market. What is the economic function of a secondary market? As seen through the eyes of a market participant, the function is to support trading—for instance, to handle the sale of stocks or the purchase of a home. A microeconomic theorist might state the thought a bit differently: the function of a secondary market is to make the asset that you want to buy or to sell more liquid.

Our references to the equity markets in this book will be almost entirely to the secondary markets where already issued shares are traded. The largest of these markets are the New York Stock Exchange (NYSE) and NASDAQ in the United States, and the London Stock Exchange and Deutsche Börse (the big German exchange) in Europe. The other major European stock exchange group, Euronext, merged with the NYSE in 2007 (the combined company is now named NYSE Euronext, Inc.).

Around the world, exchanges have changed and grown tremendously in the past quarter of a century. Technology, regulatory, and competitive developments have had enormous impacts, as have the engineering of new financial products (such as options and futures) and the growth of the customer base (investors, both retail and institutional). Much academic insight into the operations of the equity markets has developed over roughly the same period. This relatively new field in financial economics is known as *market microstructure*. Microstructure has its roots in microeconomic theory. Reciprocally, microstructure provides rich examples of how microeconomic theory can be applied to shed light on the dynamic operations of a micro market. We make extensive use of these examples in this book.

## FRICTIONLESS VERSUS NONFRICTIONLESS MARKETS

A theory is derived from a set of assumptions. The assumptions are the abstract descriptors of a reality that we wish to understand. Each assumption individually is typically unrealistic. The important thing is that, when properly brought together, the assumptions collectively can be realistic, and they can deliver the insights that we are looking for. For instance, we all know that the earth is round, not flat. However, for most applications that involve spatial measurement, Euclidian geometry is the way to go.

The specific assumptions that we make depend on the questions we are asking. Our focus in this book is on the micro markets, and we have an important choice to make: Assume that trading is a frictionless process, or allow for the existence of trading costs and other market impediments. Through Chapter 3, we assume a frictionless market in order to facilitate our discussion. It is interesting to know what Nirvana looks like. But then, starting in Chapter 4, we recognize the existence of various costs, blockages, and other frictions that, in real-world markets, attend trading.

Market breakdowns can occur. Recognizing this we can also understand that the efficiency of a market depends on its architecture. By *architecture* we are referring to the facilities offered to market participants (for example, the ability to purchase airline tickets via a computer) and the rules and regulations that govern trading (for example, the specification of standardized store hours, and the enforcement of minimum quality standards). Why do the facilities, rules, and regulations matter? The reason is that real-world markets are not frictionless environments. With regard to equity markets, for instance, translating fundamental information about a company into share prices is a complex, imperfect process in the nonfrictionless world. Simply handling orders and turning them into trades can be a big challenge. Consequently, theoretically desirable values for price and trading volume may not be attained.

Regarding the secondary markets for equity trading, we will explain how buy and sell orders are submitted to the market, interfaced with each other, and turned into trades and transaction prices. We will see that transaction costs, blockages, and other impediments give rise to coordination problems that can prevent the market from realizing theoretically desirable values for price and trading volume. In this context, we may gain an understanding of how a market's trading facilities, its rules, and its regulations affect the efficiency with which individual desires to trade get translated into realized transactions and transaction prices.

And so we devote serious attention to nonfrictionless trading. However, the frictionless environment continues to have an important role to play. The zero trading cost model is an ideal against which to assess real-world markets.

## HOW WELL DO MARKETS FUNCTION?

You cannot study microeconomic theory without developing a healthy respect for the power of a free market to deliver results that are desirable for society. This property of markets is vibrantly reflected in Adam Smith's description of the *invisible hand*:

> *Every individual necessarily labors to render the annual revenue of the society as great as he can. He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this case as in many other cases led by an* invisible hand *[emphasis added] to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectively then when he really intends to promote it.*
>
> —Adam Smith, *Wealth of Nations* (1776), p. 423

Microeconomics has been defined as the study of the allocation of scarce resources between competing ends. If markets were allowed to operate freely, would the allocation of different resources in production be socially desirable? Would the payments for the various productive factors be socially desirable? Would the allocation of resources across consumption goods be socially desirable? Achieving these results is indeed a lofty objective. It is the noble objective of a market-based economy. Can it be achieved? Does Adam Smith's invisible hand do all that Mr. Smith suggests it might? Quite strikingly, it can be demonstrated, under a simplifying set of assumptions, that these objectives would be met in a free and competitive market environment. Undoubtedly, the power of this proof accounts in large part for the basically free market stance of many microeconomic theorists.

That said, various conditions would have to be satisfied for free markets to operate with a high degree of perfection. One condition above all others has received a great deal of attention: The markets would have to be competitive. Competition is the force that drives profit-maximizing firms to achieve socially efficient results. In a simplified, competitive marketplace, price is an all-important guiding variable that competitive economic agents respond to. However, for prices to properly guide production and consumption decisions, no competitive entity can individually be big enough to have the power to influence any price. Rather, all prices must be set in the marketplace with every participant being a *price taker*, not a *price setter*.

Think of this from your own perspective. Pick a good or service that you consume. Say it is bread. You could not possibly be such a big buyer of bread that you could have the market power to influence its price. When you go to the market to buy a loaf, you are a *price taker*. You see the price and respond to it. When all economic agents are price takers, price is set in the marketplace by the underlying forces of demand and supply. We explain how this works in Chapter 3.

But what if a particular good or service is produced by just one firm? When there is just one supplier in the market, the firm is a *monopolist* (when there is just one buyer in the market, the firm is a *monopsonist*). A monopoly firm is not a price taker, it is a price setter, because it has power over price. By decreasing output, a monopoly firm can increase its price. Differently stated, if the monopoly firm sets a higher price, it is willing to accept selling less. Accordingly, monopolists do something that microeconomists are not happy about: A monopolized market compared to a competitive market supplies a smaller output and charges a higher price. This result undermines efficiency for the broader economy. It also argues for government intervention in the operations of the free market. Specifically, it leads to antitrust regulation. We consider the pros and cons of government intervention in further detail in Chapter 8.

With regard to the competitive structure of an industry, monopoly is at one end of the spectrum (there is just one firm in the industry), and *perfect competition* is at the other end (the number of firms in the industry is large enough so that each and every individual firm is insignificant in size relative to the total). Models of perfect competition further assume that the output of any one competitive firm is indistinguishable from the output of any other competitive firm in the industry (that is, one farmer's red winter wheat is indistinguishable from another farmer's red winter wheat). In perfectly competitive markets, all participants are price takers, not price setters. Do perfectly competitive markets always produce socially desirable outcomes? They do not.

There are other reasons why a free market may not operate perfectly, and these reasons apply to all markets regardless of their competitive structure. Here are some key ones (these along with a few others are discussed in further detail in Chapter 7):

**Trading costs:** Markets are not frictionless; traders incur transaction costs.

**Imperfect information:** The information that we base our decisions on and that we act on may be enormous in scope, complex, imprecise, and incomplete.

**Asymmetric information:** Some people whom we deal with (trade with and make contracts with) are better informed than we are.

**Irrational behavior:** Some people act irrationally.

**Externalities:** Some innocent third parties who do not participate in a trade may be affected by the trade (for instance, the generation of air pollution in the production of electricity).

**Moral hazard:** A person who undertakes an activity without bearing full responsibility for all possible outcomes may not take into account harm that may be done to others.

**Public goods:** Goods and services (like military defense) that people may collectively benefit from but that cannot be purchased and sold individually.

**Systemic risk:** The risk of broader, systemwide breakdown and possibly failure.

It is clear that unregulated micro markets do not operate perfectly, that clearly identifiable causes of market failure exist. What should public policy be in light of this? To what extent should government step in and attempt to rectify undesirable situations? How much faith should we have in the ability of a free market to police itself? How much trust can we place in Adam Smith's invisible hand? Under what conditions will a free market be more robust than a regulated market? Should regulatory intervention be confined to controlling dishonest behavior and other abuses of power and position, or should it also address matters involving the architectural structure of a marketplace? These public policy issues have been debated over the years. We consider some of them in Chapter 8.

Here is the conundrum: On the one hand, free market institutions are imperfect. On the other hand, government organizations are themselves imperfect, and government officials typically have only imperfect and incomplete knowledge of the situations that they are attempting to rectify. What has euphemistically been labeled *unintended consequences* can attend even well-intentioned legislation. One particularly serious consequence of misguided regulatory intervention is that it can undermine the one property of markets that democratic governments (and most microeconomists) typically value: the vibrancy of competition.

Both a free market's ability to produce desired results and the efficacy of government to mitigate market failings have been seen in a new light following the financial turmoil that erupted in 2008 and the frightening economic downturn that ensued. Our trust in the stability of the free market was violently shaken by a long concatenation of events that included the demise of Lehman Brothers (the venerable investment banking firm), the collapse of American International Group (AIG, the big insurance company), the freezing up of credit lines throughout the economy, and the bankruptcy of General Motors. The enormous, pervasive role that the financial sector plays in the economy was powerfully underscored by these recent events. Quite clearly, finance in the nonfrictionless world involves far more than simply transferring funds from net savers (households) to net investors (firms). Financial capital is a crucial productive factor for all firms, and it is a necessary tool for their risk management. We consider this in further detail in the book. We address the failings of the free market and the problems concerning governmental intervention in Chapters 7 and 8, respectively. But, as you will see, we conclude on a positive note: The power of a free market to ultimately produce desirable results should not be underestimated.

Competition is a pervasive force in our lives. As students we compete for grades, as athletes we compete for titles and trophies, as graduates we compete for jobs and better pay. Business firms compete for profits, politicians compete for votes, artists compete for recognition, and so forth and so on. If properly played out, the competition is inherently good. Competition can be an enormously constructive force that substantially improves the lives of individuals, the greatness of nations, and the overall quality of the global economy. In discussing a Chinese television series first aired in 2006, *The Rise of the Great Nations*, Fareed Zakaria stated that the basic message delivered by the series is ". . . that a nation's path to greatness lies in its economic prowess and that militarism, empire, and aggression lead to a dead end."[1] This is a hopeful message indeed. And it underscores the importance of having micro markets that are well-functioning, competitive, and vibrant.

At its core, microeconomics focuses intently on how competition channels activities and affects outcomes in micro markets (and, by extension, macro markets). Wise decisions can be made regarding government intervention in the marketplace only when the operations of a micro market are sufficiently and properly understood. This book has one overriding objective: to further that understanding for you.

---

[1] Fareed Zakaria, *The Post American World* (New York: W.W. Norton, 2008).

## A MARKET IS AN ECOLOGY

Reading about the economic forces that underlie a micro market is one thing. Actually operating in a micro market is something else. We all do this; typically any number of times a day, although we rarely give it much explicit thought. To underscore some of the important concepts presented in this book, we can together enter one market conceptually and see in more detail how it works. As noted earlier in this chapter, we will enter an equity market. We will see how prices are determined and how trading volumes are established in that market.

That is, we will see just how prices and transactions volume can be determined based on the *order flow* received by the market. We will show you how the order flow can be translated into downward sloping curves to buy shares, and into upward sloping curves to sell shares. You will see how the translation of orders into prices and trading volume depends on the architectural structure of the marketplace. You will come to appreciate that trading in the equity market involves making tactical trading decisions. You will also comprehend the costs involved in placing your orders, and will understand the cost pressures felt by a dealer firm when its inventory of shares has become unduly large.

Trading is very different from investing. Investing involves selecting the stocks that you wish to hold in your portfolio, while trading involves the implementation of your investment decisions. Trading is the activity that brings you to the marketplace. An equity market, because it is a rapidly moving, highly dynamic environment, brings to light some of the real-world characteristics of a micro market that we wish to be aware of and to understand. As described in Schwartz, Francioni, and Weber,

> . . . *once a (portfolio) decision is made and has been passed on to the trading desk, time acquires a different meaning. The clock suddenly accelerates. Prices in the marketplace can move with startling velocity in brief intervals of time. As they do, trading opportunities may pop up and quickly vanish. Your own order handling can cause a price to move away from you. Poor order placement and imperfect order timing can be very costly. Consequently, a hefty portion of the gains that an asset manager might otherwise have realized from a good investment decision can be nullified by a poor trading decision.*
>
> —*The Equity Trader Course* (John Wiley & Sons, 2006), p.2

Conclusion: when you go to an equity market to trade, it pays to do it right. Doing it right calls for having the best understanding possible of a market's dynamics, of just how the market behaves.

Doing it right calls for understanding that an equity market can be thought of as an *ecology*. Ecological systems involve the interactions between living organisms and their environment. When the pattern of interactions is reasonably stable, an environment is said to be in ecological balance. This view applies to an equity market. Traders interact with each other in an equity marketplace for a multiplicity of motives. Some are looking to buy shares while others are looking to sell. The limit orders of some participants (these are priced orders that are placed on an order book) bring liquidity to the market, while the market orders of other participants (these are orders to buy or to sell immediately at prices that have been set by other participants such as the limit order traders) take the liquidity that is there away. Some people trade for informational reasons, and others for their own personal reasons, including their own liquidity needs for buying or selling shares. Yet other participants jump in as momentum players who follow (and thereby reinforce) a trend, while others approach the market as contrarians, buying as prices drop and selling as they rise. All of the above is needed for an equity market to function well as an ecological system. The wording can be stronger: The multiplicity of events is *required* for an equity market to operate effectively. And a diversity of events drives the dynamics of other real-world micro markets as well.

## BASIC MICROECONOMIC CONCEPTS

Certain concepts are central to microeconomic theory. We turn to them now before progressing to the rest of the book.

### *Ceteris Paribus:* All Things Being Equal

We will be considering issues such as how much of a good, service, or financial asset will be demanded in the market, and how much will be supplied by the market. For now, let's focus on demand. By demand, we mean the amount of a resource that a user will actually seek to purchase (consider this *effective demand*), as distinct from the amount that the individual would ideally like to have if the resource was free (consider this *wishful demand*).

If we were to ask a portfolio manager (PM) how many shares of *xyz* stock he or she would hold in a portfolio under management, what do you think the PM would immediately ask us? One question would undoubtedly be, "What price would I have to pay to acquire the shares?" Right on! Regardless of the resource under discussion, how much of it you would seek to acquire would depend on its price. Price is a crucially important variable for all of us in general, and for a microeconomist in specific.

Our scarce resources are allocated across the broader set of goods, services, and financial assets that are available to us with regard to price. But price is not the only variable that counts. Other factors such as your income (and/or wealth), the price of other resources, and your tastes also matter. If we want to focus on the pure effect of price, we do so by "holding the other factors constant." The Latin term for this is *ceteris paribus*, which means "all else is equal." Interpret "all else" as everything else that is relevant, and "equal" as "constant."

So, how many shares of *xyz* will the PM wish to hold if *xyz* costs $20 a share? To answer this, we might also want to consider the investor's wealth or total assets under management, how risky *xyz* is, the earnings per share that the PM expects the company to achieve one year into the future, and the price of other stocks that he or she is considering holding.

Here are two more questions:

1. How many more shares of *xyz* would the investor wish to hold if the price was lower, say $19 per share?
2. What else might be different if the price is $19 a share rather than $20?

To assess the responsiveness of the PM's demand to hold shares to the share price, the answer to Question 2 should be "Nothing else is different." Recognizing this, Question 1 can be reworded: "How many more shares would the investor want to hold if the price were to change from $20 a share to $19 a share, all other relevant factors held constant (*ceteris paribus*)?"

Of course you might ask why the price per share has dropped from $20 to $19. Perhaps the reason is that the release of bad news has caused others to lower their expectations of future earnings. In this case, we would twist the Latin phrase around and say that "*ceteris* is not *paribus*." But that would be changing the question we started out asking. Let's stick with our original inquiry: If the price of shares were to change, *all else constant*, how would the demand to hold shares change? The answer to this question is the first step to obtaining the investor's *demand curve* to hold shares. We need this demand curve to understand how the price of shares is set in the market in the first place. We consider this demand curve in further detail in Chapter 3.

An investor's demand curve shows the array of alternative quantities that would be demanded at an array of alternative prices. As we move along a given demand curve, price is the only variable that is driving change in the quantity demanded—*all else is constant*. If any other demand determining variable changes (e.g., the investor's wealth, or the investor's investment or trading strategy, or his or her expectation of future returns for the stock or the returns expectation for an alternative, substitute investment), we say that the *demand curve has shifted*.

Let's repeat the thought with reference to a different resource: What might cause your demand for a product, such as pasta, to change? Answer: a change in any of pasta's demand-determining independent variables. If the variable is the price of pasta, your response to the price change (all else constant) would be a *movement along* your demand curve for pasta. If price is constant and some other demand-determining variable changes (perhaps your income, or your desire to cut down on carbs, or the price of rice), your response would be a *shift* of your demand curve for pasta.

A demand curve is a fundamental tool of economic analysis, and it plays a crucial role in our book. Whether the agent in question is an investor and we are considering the number of shares of a stock that are held by the portfolio manager, or the agent is the driver of a car and we are considering how many gallons of gas the driver on average purchases each week, we can conceptualize a price change and an agent's response to the price change while holding *all else equal* (i.e., *constant*). By this *ceteris paribus* procedure, we can better isolate the key factors that drive a market. In so doing, we will better understand how price is set in a market.

To summarize, *ceteris paribus* is a conceptual tool that we use to organize thinking about the interactions between a set of variables. In actual markets, lots of things can change at the same time. Nevertheless, we conceptually separate the net impacts of each of the causal factors (which are referred to as the *independent variables*) on the factor that we are interested in (which is referred to as the *dependent variable*). For empirical analysis, multivariate econometric tools exist that enable one to infer the net effect that each independent variable has on the particular dependent variable in question.

## Fixed versus Variable Costs and Return

Microeconomics is oriented to decision making; it deals with questions such as, if the price of one product (coffee, for instance) increases relative to the price of another product (tea, for instance), how much should I decrease my consumption of coffee relative to tea? Or, if a company's production costs go down, how much more output should that company bring to market? Or, if the earned income of electricians increases relative to the earned income of plumbers, how many more people will choose to be electricians rather than plumbers? And so forth.

When making a decision, all costs and returns related to that decision should be taken into account. In doing so, we make a very clear distinction: Fixed costs (and fixed returns) are differentiated from variable costs (and variable returns). Only the variable components are relevant for decision making.

Here is an example of a cost that, being fixed, should not be taken into account. Harrison and his girl friend Alba have just stepped out of a taxi that has taken them to a movie theater. Because of heavy traffic, the taxi ride cost \$10 more than Harrison had expected. He added that to the cost of the movie ticket and concluded that the "all-in price" was too high. Being frugal, Harrison suggested to Alba that they skip the movie and simply take a walk.

Alba, however, had something to say about this. Having taken a microeconomics course, she was able to point out the error in Harrison's thinking. "We are here already, we have already paid for the taxi, and that payment is now a fixed cost. As such, Harrison," she said, "it is a sunk cost. That cost is what it is, it does not depend on whether we see that nice movie or not. Therefore it is not relevant to our decision making. Come on, Harrison, let's buy the tickets." And so they did. And Alba was right.

As we proceed to analyze various *decisions*, it will always be variable costs only that we take into account. Fixed costs should be taken account of only when considering how well off a decision maker is (for example, the profits that a firm has realized from its production and sales are clearly affected by fixed costs).

## Long-Run versus Short-Run Analysis

Long-run versus short-run analysis refers to the length of time a decision maker has to respond to the change of an economic variable. For simplicity, we distinguish between two abstractly defined time horizons, a short-run horizon and a long-run horizon. The distinction honors the reality that the marketplace is not a frictionless environment, that economic agents cannot make adjustments instantaneously, and that the process of adjusting to change takes some time. Adjustments can be made in the short run, but they are only partial. For instance, if the price of an automobile decreases, additional cars may be sold in the short run, but sales will expand more in the long run as consumers gain further time to replace their existing vehicles and/or to buy additional ones.

In the long run, all costs and returns are fully variable (and all adjustments are complete), while in the short run some costs and returns are fixed. For instance, in the short run we assume that a company's plant size is fixed but that the firm is able to alter its number of employees. In the long run, a firm's workforce and its plant size are both fully variable.

How long is the long run? The answer is, "it depends." The long run is measured in years if we are thinking of the supply of cardiologists because a lengthy education is required to enter that profession. With regard to portfolio management, the long run will be measured in fractions of a year, and with regard to securities trading, an hourly interval might better define the long run.

## Equilibrium

Equilibrium plays a central role in economic analysis. We seek to understand how resources are allocated between competing ends so as to maximize profits for producers (suppliers) and to maximize utility for households (consumers). How are decisions made? How should the tradeoffs that are faced be resolved? When an individual firm or household has optimally formulated its decisions, it is in equilibrium. When the firms and households that in aggregate comprise an industry have achieved stable pricing, production levels, and consumption levels, the industry is in equilibrium. After an equilibrium has been attained, nothing will change until an exogenous shock occurs (for instance, a change in technology, or tastes, or the availability of a productive resource, or the price of a related good or factor of production, or a regulatory intervention).

We consider equilibrium for individual firms and households, and for individual industries. We consider equilibrium in the short run and in the long run. We focus on *static equilibrium* (outcomes that participants attain individually, and that markets in aggregate are pushed toward by economic forces). We consider *comparative statics* (a contrast of two different static equilibria that is attributable to a difference in the value of an independent variable). And we occasionally have a word to say about *dynamic equilibrium* (which addresses the adjustment that occurs over time when a decision maker or market is not currently in static equilibrium). Our analysis of equilibrium (static, comparative statics, and dynamic) pays particular attention to the structure of relative prices. *Relative prices* is the hallmark of microeconomic analysis. Macroeconomic analysis, in contrast, focuses on variables such as aggregate income levels, economic growth, and change in the aggregate level of prices (i.e., inflation).

## Marginal Analysis

Marginal analysis can be viewed as an extension of the thought that only variable costs matter. Marginal analysis means that we consider how costs and revenues change *on the margin* as we pursue an activity more or less intensively. Because fixed costs are fixed, the marginal costs associated with fixed costs are zero. Variable costs do change, however, which is why they are called variable. Marginal costs are associated with variable costs. Here is an example.

Harrison, a serious runner, is generally on the road four times a week, averaging five miles a run. One day he set out on an eight-mile run. The first three miles were fine as he warmed up and started cruising. The next four were increasingly challenging and, finally, mile eight had to be toughed out. At the end of it, Harrison stopped. How would a microeconomist describe the pattern of each successive mile becoming increasingly more demanding? Viewing each mile marker as a one-mile increment, the pattern is one of increasing incremental cost (that is, the incremental difficulty of running from mile 7 to mile 8 was greater than the incremental difficulty of running from mile 6 to mile 7, which in turn was greater than the incremental difficulty of running from mile 5 to mile 6, and so on).

How would an economist explain Harrison having stopped at the end of mile eight? Through mile eight, the benefit of running each additional mile must have exceeded the incremental cost of doing so. But the incremental cost was increasing, and the net benefit was decreasing with each passing mile. If Harrison had run a ninth mile, the incremental cost would have risen to the point where the net incremental benefit had turned negative. Harrison undoubtedly applied this incremental analysis (if only intuitively) to decide the length of his run.

But our formidable runner did not have to make his decision in terms of one-mile increments. If Harrison was running on a quarter-mile track, he would likely have used quarter-mile increments. But a microeconomist would shrink the measurement increment even further. Let's go all the way down—let's shrink the increment to an infinitesimal amount. Then the microeconomist would not use the term *increment*, but would say *marginal*. The decision of how far to run can be made with precision when *marginal analysis* is used: Harrison would stop precisely at the point where the marginal cost rose to a level that set the net marginal benefit of continuing equal to zero.

Here is another example of marginal analysis. Alba has a portfolio that is currently valued at $100,000. She is thinking about how she would feel if, in one year, her portfolio is valued at $90,000 (negative growth), or at $100,000 (zero growth), or at $110,000 (good growth), or at $120,000 (strong growth). Alba might state her feelings as, respectively, total disgust, disappointment, satisfaction, and total delight.

A microeconomist would depict her emotions somewhat differently. Each outcome from the lowest ($90,000) to the highest ($120,000) would be assigned a *utility value* (*utility* is the term we use for *pleasure*). Clearly, the higher the outcome, the greater is Alba's utility. Contrast two of these increments: (a) the differential utility achieved by being at $100,000 instead of $90,000, and (b) the differential utility achieved by being at $120,000 instead of $110,000. That is, how does the incremental utility gained by going from total disgust to disappointment compare in magnitude with the incremental utility gained by going from satisfaction to total delight? Which utility increment do you think would be greater? That is, which of these two jumps in portfolio performance would add more to Alba's utility index?

Economists typically assume that successive, equal-dollar increments (in this case, $10,000 increments) yield diminishing successive increments of utility. So let's presume that Alba's utility increases with dollars, but at a decreasing rate. This being the case, increment (a) is greater than increment (b). Next, let's shrink the dollar increments in question from $10,000 additions to infinitesimal additions. When we do so, we can refer to Alba's *utility function for dollars*, and to Alba's *marginal utility* of dollars. The marginal utility of dollars is the amount by which an infinitesimal increase in dollars increases her utility. Because Alba's utility is increasing with dollars at a decreasing rate, her marginal utility of dollars falls (but remains positive) as the total number of dollars that she has increases.

We can define Alba's utility function with respect to either her income or to her wealth. Either way, we see that Alba has diminishing marginal utility of income (or wealth). As we explain in Chapter 2, this property of diminishing marginal utility of income (or wealth) has important implications for Alba's willingness to accept risk.

Here are two more examples of how adjustments can be made *on the margin*.

1. An increase of wages in industry A relative to industry B leads workers to seek employment in industry A. But not all workers will feel the need to respond; only some people may change their employment. Those who do so are the ones who, either because of geographical proximity, their skill sets, and/or their personal preferences, find it most beneficial to move from one industry to the other. These are the people who are "on the margin." Change "on the margin" is all that is needed for a market to equilibrate following a price adjustment.
2. Shares of *xyz* stock are trading at $32 before bullish news comes out and the share price increases to $34. The company has 2 billion shares outstanding, but the price change occurs with only 15 million shares changing hands. Consequently, $4 billion of market capitalization was created by a dollar-trading volume of roughly $480 million. How could

this happen? The reason is that price is set in the marketplace by the subset of participants who meet in the market and trade. The subset is comprised of the *marginal participants* (those who for whatever personal reason, perhaps their assessments of information or individual cash flow needs, are most apt to buy or to sell shares). Again, the message to be received is that everyone need not respond to news for the new information to be reflected in price. As in all markets, price is determined by activity that takes place *on the margin.*

## Elasticity

In our discussion of *ceteris paribus*, we have considered how a dependent variable (e.g., the quantity of a good demanded) is related to a set of independent variables (e.g., the price of the good in question, the price of other related goods, and income). Part of understanding this relationship involves knowing the *responsiveness* of the dependent variable to each of the independent variables. *Elasticity* is our measure of that responsiveness.

One way of measuring the responsiveness of some variable $y$ to some other variable $x$ is simply to divide the change in $y$ (call it $\Delta y$) to the change in $x$ (call it $\Delta x$). And so, we could let $\Delta y/\Delta x$ be the measure, where $\Delta$ signifies a discrete (incremental) change. For one reason in particular, however, this is not a good measure: The ratio is affected by the units of measurement used for both $y$ and $x$. To understand the problem that this can lead to, first let $y$ be the number of gallons of gas that a "representative" consumer purchases per week, and let $x$ be the price per gallon of gas. Perhaps we observe that a 10¢ increase in price results in a two-gallon decrease in demand (in other words, that $\Delta y/\Delta x =$ –2 gallons/10¢, where the minus sign indicates that the consumption of gas *decreases* when the price of gas rises at the pump).

Now let $y$ be the number of pounds of tomatoes that the "representative" consumer purchases per week, let $x$ be the price per pound of tomatoes, and assume that the price of tomatoes goes up 10¢. Perhaps we observe a value $\Delta y/\Delta x = -.5$ pounds/10¢. Question: The price of both items, gas and tomatoes, has increased by 10¢. Which of the two items is more responsive to the price change, gasoline or tomatoes?

Who can say? The problem is that a gallon of gas cannot be compared to a pound of tomatoes. There is, however, a simple way to make the two responses comparable: contrast the *percentage change* in $y$ with the *percentage change* in $x$. By this adjustment, we eliminate the labels associated with $y$ and $x$ (e.g., *gallons* of gas and *pounds* of tomatoes). The simplest way to do this adjustment is to use marginal analysis instead of incremental analysis, and to take our measure of responsiveness to be

$$\eta_{y \cdot x} = \mathrm{d}y/y \div \mathrm{d}x/x = (\mathrm{d}y/\mathrm{d}x)(x/y)$$

where $\eta_{y \cdot x}$ denotes the elasticity of $y$ with respect to $x$, and d$y$ and d$x$ are infinitesimal changes in $y$ and $x$, respectively.

We are taking $y$ to stand for quantity demanded. If $x$ is the per unit price of $y$, $\eta_{y \cdot x}$ is the own-price elasticity of demand. If $x$ is the representative consumer's income, $\eta_{y \cdot x}$ is the income elasticity of demand. And so on.

This is enough for now on elasticity. We return to this topic in Chapter 3.

## Maximum, Minimum, and Optimum

Some things we wish to maximize, other things we wish to minimize, and for most things we want to select an optimal amount. For instance, we all want to maximize happiness, but, for a multiplicity of decisions that we make, the objective is to optimize. Do you like sugar in your coffee? If so, how much? Right now, in the context of the current discussion, "an optimal amount" is the perfect answer.

Consider the construction of an equity portfolio. Three characteristics of a portfolio are of overriding importance to an investor: its risk, its expected return, and its liquidity. For now, let's focus on just the first two and on the relationship between them. The higher the expected return that the portfolio yields, the happier we are. On the other hand, the riskier the portfolio, the more pain we feel. For any level of risk we certainly want to maximize the expected return. Similarly, for any level of expected return, we wish to minimize the level of risk. But with risk and return both being variables, our objective is to achieve an optimal combination of the two. Fortunately, we can control the risk and return characteristics of a portfolio by combining the individual financial assets that comprise it in appropriate proportions. Let's pursue this further.

When financial assets are properly priced, a clear trade-off exists between risk and return: Higher risk assets (both single stock positions and portfolio combinations) yield higher expected returns, and stocks and portfolios that offer lower expected returns are less risky. Thus investors face a trade-off: They can realize higher expected returns only by accepting more risk, or they can reduce their risk exposure only at the price of accepting lower expected returns. Facing this tradeoff, the investor should select a portfolio that offers an *optimal* combination of risk and expected return. The desired portfolio is optimal in that obtaining it maximizes the investor's utility, which is a function of wealth (we explain more about this in Chapter 2).

"Optimal" is like Goldilock's preferred porridge: It could be hotter, it could be cooler, but it is just right. A "just right" amount is determined with reference to Goldilock's single, ultimate objective: maximize her utility. In microeconomic theory, the single, ultimate objective for a consumer is just that: maximize utility. For a firm the goal is to maximize profits. In pursuit of these objectives, the decision makers make optimal choices from the menu of alternatives available to them.

## FRICTIONLESS MARKETS, THE GULF STREAM, AND A DAY AT THE RACES

Two contrasts of considerable importance have been set forth in this chapter: (1) theory versus reality, and (2) frictionless versus nonfrictionless markets. With respect to the first contrast, we care about reality, but we use theoretical analysis to gain insight into the complex workings of real-world micro markets. With respect to the second contrast, real-world markets are replete with trading frictions, but we can assume these impediments away in order to simplify and thereby expedite much of our theoretical analysis. But if your study of microeconomics is confined to frictionless markets only, you might be left thinking that the subject is simply too unrealistic to be useful. Although we would disagree with your conclusion, we would certainly understand your thinking that way. But rest assured, starting in Chapter 4, we very definitely pay attention to the realities of the nonfrictionless marketplace.

However, do not be too fast to dismiss the frictionless environment. It serves two purposes, and it serves them well. For one thing, friction or no friction, strong underlying forces drive the micro markets, and it is important to understand them in their pure form. You might better appreciate this with the aid of an analogy. The Gulf Stream is a warm, powerful ocean current that has its origin in the Gulf of Mexico, after which it follows the eastern coastlines of the United States and Newfoundland before traversing the Atlantic to the shores of Europe. The ocean that it crosses is often beset by winds, heavy precipitation, and waves that in a storm can be fierce. If you care about things such as average temperatures and weather patterns on both sides of the Atlantic, you must have knoweldge of the Gulf Stream. But if you are a ship's captain, you undoubtedly also care about the wind, precipitation, rip tides, and waves on the surface of the sea. To the ship's captain, the Gulf Stream may be theory, and the surface conditions a reality. But they both matter. A wise captain should, and no doubt does, know and care about both.

Depicting the frictionless market also serves a second purpose. It is an ideal toward which we might forever strive but in reality never achieve. Superior technology, more efficient market architecture, better informed and more sophisticated participants may all help but, with regard to each of these, perfection is forever impossible to achieve. The concept of a frictionless market is like the mechanized rabbit that greyhounds chase at a race. The rabbit is always in front of the charging dogs, the rabbit is the hounds' goal, but the rabbit can never be caught. Yet it serves a very important purpose.

## CHAPTER SUMMARY

In this introductory chapter we have established a number of key concepts that will be important for our market structure presentation of microeconomics. Here are some highlights of the chapter.

1. Microeconomics is the study of how economic units determine what and how many goods and services to consume (households) and produce (firms), and what and how much of the factors of production (for example, labor and capital) to supply (households) and to use in production (firms).
2. Microeconomic theory formalizes basic principles concerning how best to resolve the tradeoffs involved in consumption, production, and other allocation decisions. It does so by tying assumptions together to form an economic model. The assumptions individually need not be perfect reflections of real-world economic activity. Theory, by its nature, is abstract, while real-world markets can be very intricate. The hallmark of a good theory is that, although abstract, it provides insights that enable us to understand better the operations of the real-world micro markets.
3. Market structure refers to the architectural realities of actual marketplaces. These encompass the institutional features that pertain to market design and operations.
4. The power of microeconomic theory can be better understood by applying it to an actual, real-world market. Throughout the book, as we extend our theoretical discussion to a nonfrictionless market characterized by imperfect information, trading costs, and blockages, we focus on the equity markets. Specifically, we consider the secondary markets where already-issued equity shares are traded (not the primary markets where new shares are issued).

5. Equity market microstructure, a relatively new field in financial economics, considers the processes by which orders are submitted to a nonfrictionless marketplace, are handled in the marketplace, and are turned into trades and transaction prices.
6. Market participants (buyers/consumers and sellers/producers) interact to set the prices of goods and services in the context of specific institutional environments. Through price setting, markets allocate scarce resources to participants according to the relative prices that they face, their incomes (or wealth positions), and their tastes for the various goods and services as described by their utility functions (and also by technology which is described by production functions, a topic that we discuss in Chapter 5).
7. A vast array of different markets exists, including commodities (such as coal, crude oil, wheat, and wood), labor (full-time and part-time), managerial resources, and capital (physical, human, and financial.) Equities markets (one of the markets for financial capital) are an excellent target for market microstructure analysis.
8. We have addressed the question, how well do markets function? On the one hand, there is the positive force exerted by Adam Smith's invisible hand. On the other hand, the forces of competition that the invisible hand represents are impeded by some agents having market power (at the extreme, monopoly power), and by various other factors including trading costs, imperfect information, and systemic risk.
9. To understand better just how prices and transactions volumes are determined in an equity market based on the order flow received by the market, and to appreciate that trading in an equity market involves making tactical decisions, we have noted that the market can be thought of as an ecology. Traders interact with each other for a multiplicity of motives, and a multiplicity of motives is, in fact, required for an equity market to operate effectively.
10. We have presented seven key concepts: *ceteris paribus*, fixed costs versus variable costs and return, long-run versus short-run analysis, equilibrium, marginal analysis, elasticity, and maximum versus minimum versus optimum. These concepts underlie much of the discussion in this book. It is important to have a good grasp of them. If you feel comfortable with these concepts, you will better appreciate your study of this market structure presentation of microeconomics.

## QUESTIONS TO THINK ABOUT

1. Select a micro market that you have some familiarity with and/or are curious about. Possibilities could include movie theaters, the oil market, restaurants, or the textbook market. Write down some distinctive, distinguishing characteristics of the market that you have chosen. In some markets, prices are set by the abstract forces of demand and supply, in some markets prices are administered, and in others they are negotiated. How are prices set in the market that you have selected? Also note one or two questions concerning the operations of that micro market that might intrigue you.
2. A number of assumptions underlie our microeconomic models. These include: The marketplace is a frictionless environment, participants have perfect information, the marginal utility from consuming individual goods and services is diminishing, no individual participant in a market is large enough to affect a market determined price. How do you feel about these assumptions? Could a set of assumptions that are individually unrealistic, collectively have descriptive/predictive power? What is the alternative if you do not make some of these assumptions?
3. Liquidity is a tricky concept. You might know it when you see it, but how would you define it? Give this some thought and then pick two markets, one that you consider relatively liquid and one that you consider relatively illiquid. Contrast the two markets in terms of your definition of liquidity. How might the relative absence of liquidity in one of these markets affect your willingness to go to that market to trade?
4. Consider your decision of whether or not to do something such as make a substantial purchase (perhaps an automobile) or take a lengthy vacation trip. What are some of the costs (frictions) involved in implementing a decision to go ahead with the purchase or to take the trip? Examples of an implementation cost include the hassle of moving from one house to another, of going through airport security before boarding a flight, of getting a bank loan before buying a car, or of incurring transaction costs when buying/selling a financial instrument. Might you be dissuaded from proceeding by the implementation costs alone? If your answer to any of the above questions is yes, can you see that trading frictions can change market outcomes?

(*Continued*)

(*Continued*)

5. What do you think about the power of Adam Smith's invisible hand to produce desirable micro-market outcomes? We have presented eight major reasons why a real-world, competitive market may not deliver theoretically desirable outcomes. Which of these reasons do you consider the most important? Can you provide specific examples of some of the reasons in action?
6. Select two different goods or services, one that you expect would be relatively price-elastic and one that you expect would be relatively price-inelastic. Explain the basis for your elasticity expectations.

## TraderEx SIMULATION: LET'S LOOK AT A MARKET

The chapter started with the statement that our subject, microeconomics, is about the functioning of a market. Let's take a look at a specific market. With the use of a simulation model, we can focus in on one micro market that is both of critical importance to the macro economy and is particularly intriguing: a market where already issued equity shares trade. We can do this with the use of the TraderEx simulation software that is available in *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course*. TraderEx is an interactive computer simulation that is designed to increase your understanding of how real-world equity markets operate, and to provide you with hands-on experience in making tactical trading decisions.

*Mastering the Art of Equity Trading* provides further information about the structure of the equity markets and about how to use the software. Specific exercises are also provided. We invite you to use *Mastering the Art of Equity Trading* in the end-of-chapter material in the other chapters of this book. At this time, we turn to the simulation software to get a first glimpse of how an equity market can function.

The order-driven market is a good environment to start with. By order-driven, we mean that some market participants who are looking to buy or to sell shares for their own portfolio reasons have posted buy or sell orders. These posted orders are referred to as limit orders, and we discuss them in greater detail in Chapter 4. Because the limit orders posted by some participants establish the prices that other participants can trade at in the market, the environment is called "order-driven." On the TraderEx screen, you will see limit orders that have been placed on the *limit order book*. A *limit order* means that a price limit has been placed on the order, and the participant is willing to wait for it in order to execute the order (and is willing to take the risk that the price limit will not be met). For a buy limit order, the price limit is the highest price the buyer is willing to pay. For a sell limit order, the price limit is the lowest price the seller is willing to receive. Do you want to buy? Enter a buy limit at your price. Want to sell? Enter a sell limit at your price. But there is also another way that you can buy and sell. You can enter a market order and trigger a trade against an already-posted limit order. The order entry box for doing this is at the top of the limit order book. If you buy "at market," you buy at the lowest posted offer (assuming that the book has sufficient depth). If you sell "at market," you sell at the highest posted bid (again, assuming that the book has sufficient depth).

Look at TraderEx's limit order book and start to become familiar with it. Get a feel for how buy and sell pressures interact in a marketplace to produce transaction prices and trading volume, and think of these buy and sell pressures as *demand and supply*. In our basic models, demand and supply jointly determine price and quantity in a marketplace. Think about how the orders you see on the limit order book collectively reflect the demand and supply propensities that have been openly expressed in the market.

But what about market orders? Until the market orders have been entered, are they *unexpressed, latent* demand and supply propensities? Hmm, this is something to think about.

CHAPTER 2

# The Consumer Choice Model

## What Would You Really Like to Do?

Micro markets are driven by two powerful forces: the demand for goods, services, and assets, and the supply of goods, services, and assets. Taking a high level, aerial view of the micro markets, we can see how these two forces play out around the following:

- **Two types of goods and services:** those that are used in consumption, and those that are used in production.
- **Three types of productive factors:** land, labor, and capital.[1]
- **Three types of capital:** physical (e.g., plant and equipment), human (e.g., education and health), and financial (e.g., equity and fixed income securities).
- **Two types of participants:** households and firms. Households demand consumer goods and services, and they supply labor services and financial capital to firms. Firms supply consumption goods and services to households. They also supply the three types of capital, along with intermediate goods and services to each other, and they demand the three factors of production.

In this chapter we focus on one part of the categorization: The demand of households for consumption goods and services, and their demand to hold shares of a risky financial asset. A household's demand to hold shares of a risky financial asset is one side of the coin; the flip side is its willingness to supply financial capital to firms. The task of the current chapter is to establish the analytic framework for analyzing demand. To this end, we first deal with two generic consumption goods (*X* and *Y*), and then apply the analysis to two trade-offs that are more concrete: (1) the trade-off between present and future consumption, and (2) the trade-off between two attributes of a financial asset, risk and return. The resolution of each of the three trade-offs is a static equilibrium solution. We end the chapter by showing the comparative statics that apply when two different values can be assigned to the price variable. The comparative statics position us well to obtain the demand curves for goods, services, and financial assets, a topic that we turn to in Chapter 3.

---

[1] Although this is not conventional, in Chapter 7 we also refer to information and to financial capital as factors of production.

## THE CONSUMER CHOICE MODEL AND THE *X, Y* TRADE-OFF

The representative household (consumer) has a single goal: *maximize his/her utility* (pleasure) that is derived from his/her income. If you get pleasure from your work (as many of us do) you can also include this nonmonetary income in your utility calculation. If you have asset holdings and anticipate receiving a stream of income payments in the future (for example from earnings, interest payments, dividends, and capital gains), you can think in terms of your wealth rather than your income. But to keep things simple, we will for the most part refer only to income, not wealth.

The representative household maximizes utility by *optimally allocating* its income over the consumption alternatives available to it. The optimal allocation is determined with reference to the household's income and the market prices of the available goods. In considering the price of acquiring any good or service, or of pursuing any activity, only variable costs are taken into account. To obtain the household's demand curve for a specific good or service, we let its price change while holding all other demand-determining variables constant (our *ceteris paribus* procedure).

In this section of the chapter we establish the underlying framework used to analyze consumer choice. We then apply the model to two major decisions made by a household: (1) how much to spend on consumption items today and how much to save and thereby spend on consumption items in the future, and (2) how much of a financial portfolio to invest in risky assets, and how much to place in a riskless asset so as to achieve an optimal risk and return configuration.

The consumer choice model in the absence of risk and uncertainty has one goal: to maximize utility. With risk, the goal becomes the maximization of *expected utility*. Maximization is achieved with reference to two major constructs: (1) the budget constraint that demarks the possibilities available to the decision maker, and (2) the decision maker's utility function as described by a family of indifference curves. We will look at each of these in this section with respect to two undefined goods, $X$ and $Y$. Applying the analysis to a trade-off between two real-world goods that you care about calls for further definition, as you will see in the two sections that follow.

## The Budget Constraint

To determine our representative consumer's selection of an optimal combination of $X$ and $Y$, a good first step to take is to identify exactly what specific combinations (baskets) of the two goods are available, given the prices of each and the consumer's income. To do this, we use the following notation:

$\boldsymbol{M}$ = the consumer's total income received per year

$\boldsymbol{X}$ = the consumer's consumption of good $X$ per year

$\boldsymbol{Y}$ = the consumer's consumption of good $Y$ per year

$\boldsymbol{P_X}$ = price per unit of $X$

$\boldsymbol{P_Y}$ = price per unit of $Y$

Note that income is expressed per unit of time (we arbitrarily selected an annual rate), and that $X$ and $Y$ are expressed as quantities consumed per unit of time (we also selected an annual rate for the quantities so as to be consistent with income being measured as a rate per annum). Variables expressed as rates per unit of time are *flow variables*. We might, for instance, say that over the last year somebody earned $80,000. Corporate income statements also report flow variables. As such, you will always see the period of time that a corporate income statement covers (e.g., the third quarter of a particular year).

Variables measured at a specific point in time are *stock variables*. The number of items in inventory is a stock variable, not a flow variable. Similarly, wealth is a stock variable. As of a specific date, we might say that a person's wealth is $50,000. Corporate balance sheets report stock variables. As such, you will always see the exact date that the measurement of assets, liabilities, and net worth applies to.

In much of this book we deal with flow variables. When we do not, it should be clear or we will say so. Here is our simple setting. Suppose that, over the course of a year, an individual consumes $X$ units of the first good at a price of $P_X$ per unit, and $Y$ units of the second good at a price of $P_Y$ per unit. Setting the sum of these two expenditures equal to the individual's annual income, we have,

$$M = P_X X + P_Y Y \tag{2.1}$$

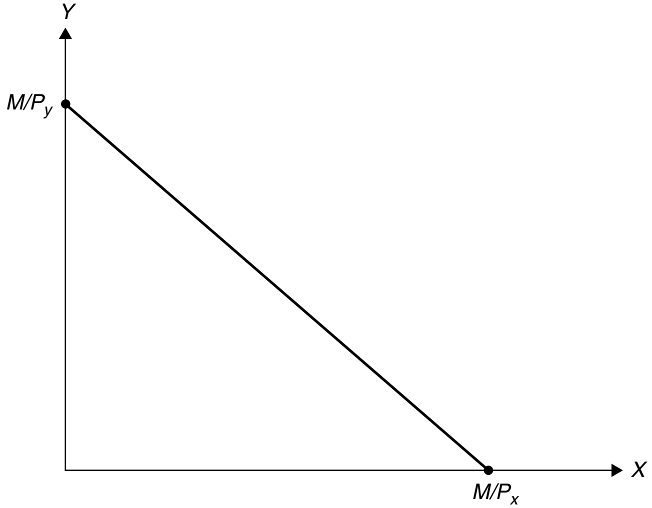


**EXHIBIT 2.1** A Consumer's Budget Constraint

The equation can be rewritten to show the maximum amount of $Y$ that can be consumed for any amount of $X$ consumed:

$$Y = \frac{M}{P_Y} - \left(\frac{P_X}{P_Y}\right)X \tag{2.2}$$

This equation is the budget constraint. We present it graphically in Exhibit 2.1. The intercepts are $M/P_Y$ (on the vertical axis) and $M/P_X$ (on the horizontal axis). From both equation 2.2 and the intercept terms in Exhibit 1.1, we see that the slope of the budget constraint is

$$\frac{dY}{dX} = -\frac{P_X}{P_Y}$$

What this tells us is that the price of $X$ relative to the price of $Y$ is the relevant variable, not the price of $X$ or of $Y$ by themselves (unless, of course, we refer to the price of $X$ *given* the price of $Y$ (which means, given that the price of $Y$ is constant). In fact, if $M$, $P_X$, and $P_Y$ were to each change by the same proportionate amount, $k$, the budget constraint would be totally unaffected. That is,

$$Y = \frac{kM}{kP_Y} - \left(\frac{kP_X}{kP_Y}\right)X = \frac{M}{P_Y} - \left(\frac{P_X}{P_Y}\right)X \tag{2.3}$$

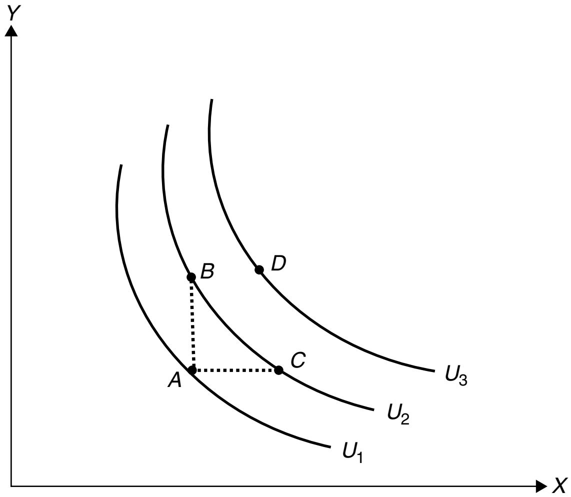


**EXHIBIT 2.2** A Mapping of a Consumer's Indifference Curves

This is a hallmark property of microeconomic theory: Absolute magnitudes generally do not matter; relative magnitudes do.

## Indifference Curves

The tastes of our representative consumer are described by his or her *utility function* for goods, services, and assets. Think of a utility function as a mountain that you can climb higher on by receiving more of good $X$, or more of good $Y$, or more of both. In Exhibit 2.2 we have presented a partial description of a representative consumer's utility mountain. The exhibit displays three curves that microeconomists call *indifference curves*. Be careful not to confuse an indifference curve with the utility function.

The utility function is the underlying mathematical statement that describes the entire utility mountain. A family of indifference curves can be obtained from the utility function. Each individual indifference curve may be thought of as a contour line on a topographical map. Just as a contour line connects all points on a mountain that are the same distance above sea level, an indifference curve connects all points on the utility mountain that yield the same utility.

There are two important differences between the mappings of these mountains, however:

1. For any contour line, the number of feet above sea level is a *cardinal number* (e.g., 100 feet above, 200 feet above, etc.). In contrast, the label

on an indifference curve is an *ordinal number* (e.g., first, second, etc., or, equivalently, least preferred, second-least preferred, etc.).

2. Actual mountains have finite peaks, while a utility mountain can have an infinite peak. Of course, consuming enough butter, eggs, or beer will leave us satiated, but this need not be the case if we define our goods broadly enough. No matter how wealthy a person might be, he or she need never be satiated by a sufficiently broadly defined good such as current consumption or future consumption. We presume that more of these goods will always make our representative consumer better off regardless of how much of them he or she can currently command.[2]

The level of utility reached by any combination of the two goods is unique to the individual whose tastes are being depicted. As we move from point *A* to point *B* or *C*, and then on to point *D* in Exhibit 2.2, the consumer gains utility. The three curves are labeled according to the level of utility received, with $U_1$ being less than $U_2$, which is less than $U_3$. Just how much utility is gained as we move from one indifference curve to another (e.g., $U_1$ to $U_2$), depends on the person's utility function and, as we have said, the increment is not subject to cardinal measurement.

With cardinal numbers, the difference between two values has meaning. For instance, picture several baskets of apples lined up according to size, from the largest basket to the smallest. You could designate the largest as first, the second-largest as second, and so on. "First" and "second" are ordinal numbers. How much bigger is the first basket than the second? Ordinal numbers do not say. To know this, we have to assess size using cardinal numbers. Perhaps the first basket contains 25 apples and the second contains 20 apples: "25" and "20" are cardinal numbers and the difference between them is meaningful. The first basket contains five more apples than the second.

The shape of each indifference curve is unique to an individual, to his or her tastes. Consider points *B* and *C* on the indifference curve labeled $U_2$. Being on the same indifference curve, the quantities of good *X* and good *Y* described by the two points yield the same utility for the particular consumer whose tastes the mapping describes. A move from *B* to *C* involves giving up some *Y* so as to enjoy some more *X*. Because a move from point *B* to point *C* leaves the consumer on the same indifference curve, the amount of utility lost by consuming less *Y* (the loss is $\Delta Y < 0$ times the incremental utility of *Y*) must exactly equal the amount of utility gained by consuming more *X* (the gain is $\Delta X > 0$ times the incremental utility of *X*). To repeat:

---

[2] The section "Questions to Think About" at the end of Chapter 8 further pursues the concept of utility mountains having infinite peaks.

Going from *B* to *C* generates a utility loss (the incremental utility of *Y* times the negative change in *Y*) and it also generates a utility gain (the incremental utility of *X* times the positive change in *X*); and since *B* and *C* are on the same indifference curve, these two changes must be equal.

But incremental change is tedious to work with. Let points *B* and *C* be infinitesimally close to each other (i.e., make $\Delta Y$ and $\Delta X$ infinitesimal). Now the incremental utilities become marginal utilities and we can write

$$MU_Y dY + MU_X dX = 0 \tag{2.4}$$

where $MU_Y$, $MU_X$ are the marginal utilities of *X* and *Y*, respectively. From equation (2.4) we have,

$$\frac{dY}{dX} = -\frac{MU_X}{MU_Y} \tag{2.5}$$

where $dY/dX$ is the slope of the indifference curve. The ratio of marginal utilities on the right-hand side of equation (2.5) is referred to as a *marginal rate of substitution* (*MRS*). In this case, it is the marginal rate of substitution of *X* for *Y* ($MRS_{X \cdot Y}$). Just as the slope of the budget constraint tells us the rate at which we *can substitute* one good for the other good in the marketplace, the slope of the indifference curve tells us the rate at which we *would have to substitute* one good for the other good for our total utility to remain constant.

What do you notice about the slope of the indifference curve that is different from the slope of the budget constraint? As you can see in Exhibit 2.1, the slope of the budget constraint is a constant that reflects the nonvarying market price of *X* relative to the market price of *Y*. As you can see in Exhibit 2.2, the slope of the indifference curve, on the other hand, becomes a smaller negative number (smaller in absolute value) as we move from left to right, continually substituting more *X* for *Y*.

## Optimal Resolution of the *X*, *Y* Trade-Off

By bringing the budget constraint and the indifference mapping together, we can understand how our representative consumer maximizes his or her utility by selecting an optimal combination of *X* and *Y*. The solution is shown in Exhibit 2.3. It is given by the point labeled *C* where an indifference curve is tangent to the budget constraint. The indifference curve that is tangent to the constraint is the highest indifference curve that can be reached.

On the vertical axis, the point labeled $Y^*$ is the optimal amount of *Y* to consume, and on the horizontal axis the point labeled $X^*$ is the optimal

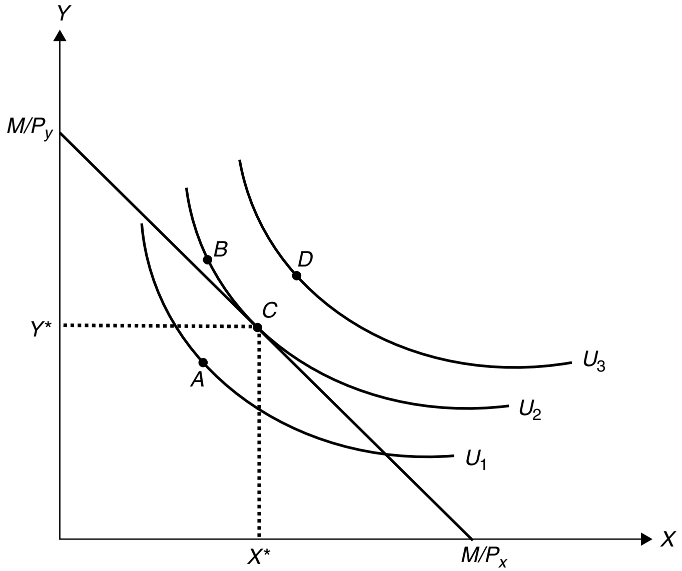


**EXHIBIT 2.3** Determination of an Individual's Optimal Consumption of *X* and *Y*

amount of *X* to consume. Note that basket *A* in the diagram lies on a lower indifference curve than does basket *C*, and that baskets *B* and *D* are not attainable. We can go further with this. Generalizing, no other attainable point in the graph yields as much utility as basket *C* and, alternatively stated, no basket that yields more utility than *C* is obtainable (for instance, basket *D*). Therefore, *C* is the very best basket for our now contented decision maker to select.

The tangency condition that is satisfied at point *C* means that the slope of the budget constraint (which reflects the rate at which *Y* and *X* *can be exchanged* for one another in the marketplace) equals the slope of the indifference curve (which reflects the rate at which the decision maker is *willing to exchange* one good for the other on the margin while keeping total utility constant). That is, the price of *X* relative to the price of *Y* equals the representative consumer's marginal rate of substitution for the two goods:

$$\frac{MU_x}{MU_y} = \frac{P_x}{P_y}$$

And so, by optimally allocating its income between the two goods, the household harmonizes its taste (on the margin, of course) with the market determined price of *X* relative to the market determined price of *Y*.

## THE CONSUMER CHOICE MODEL AND THE CURRENT CONSUMPTION, FUTURE CONSUMPTION TRADE-OFF

As we have seen with respect to the *X*, *Y* trade-off, the consumer choice model shows the optimal allocation of the decision maker's income (or wealth) among goods and services (or between assets). The model takes into account the consumer's tastes, income or wealth, and the prices of all of the goods, services, and assets that he or she is choosing between.

The model is simplified (and graphical presentation enabled) by restricting the choice to just two consumption items. We can do this with considerable realism by aggregating all goods into two baskets. In the preceding section, the two undefined goods were *X* and *Y*. In this section, we get more real. We consider the choice between current consumption (one bundle of goods and services) and future consumption (another bundle of goods and services).

At this stage of our analysis, we continue to keep our consumer choice model simple by operating in a frictionless environment characterized by zero risk and perfect certainty (knowledge) about all prices, product characteristics, and tastes. We deal further with transaction costs in Chapter 4, and with risk and uncertainty later in this chapter. But for now, do not allow for any transaction costs, risks, or uncertainty. And so our representative household knows what its current and future income will be over a known planning horizon. If that planning horizon includes multiple years, we can retain the simplicity of a two-good environment by letting next year's income and consumption represent the entire stream of future income receipts and consumption expenditures. That is, to understand the principles involved, we need not concern ourselves with the further allocation of income between year 2 consumption, year 3 consumption, and so on.

Even though we are continuing to treat a simplified environment, you will see that the discussion becomes a bit more complex as we set aside the two undefined goods, *X* and *Y*, and deal with two goods, current consumption and future consumption, that our representative consumer actually cares about. Two specific issues have now to be addressed:

1. Specification of the relative price variable (the price of present consumption relative to future consumption).
2. The location of the budget constraint.

But we are getting ahead of our story.

## Income and Consumption Patterns

Income and consumption are flow variables. Income is typically received periodically throughout the year (e.g., salaries might be paid biweekly and dividends are usually paid quarterly), while consumption expenditures are generally made on a daily basis. Let's continue to simplify, however. With no loss of generality, assume that all income receipts and expenditures are made at specific points in time. Let 0 denote the current point in time. We can think in terms of an annual increment and use the designations "point in time 0," "year 0," and "the current year" interchangeably. Moving forward to the next point in time, we similarly use the designations "point in time 1," "year 1," and "next year" interchangeably.

Let the parameters $y_0$ and $y_1$ be the household's income for the current year and the next year, respectively, and let the variables $c_0$ and $c_1$ be the household's consumption for the current year and the next year, respectively. Without being able to save, lend, or borrow, an intertemporal reallocation of resources would not be possible. Consumption in each year would simply equal that year's income (i.e., $c_0 = y_0$ and $c_1 = y_1$), and there would be nothing more to say. The intrigue begins only when we introduce the ability to save, lend, and borrow. We can do this in a very simple, insightful, and albeit unrealistic way: *assume unlimited borrowing and lending at a constant interest rate, r.*

By saving and lending some current income at point in time 0, our decision maker can consume more than otherwise one year later (at point in time 1). Or by borrowing against year 1 income, our decision maker can consume more in the current year (at point in time 0). Here is how it looks for the decision to save:

$$c_1 - y_1 = (y_0 - c_0)(1 + r) > 0 \tag{2.6}$$

The term on the left-hand side of equation (2.6) is the positive increment to time 1 consumption that is attributable to time 0 saving, and the first parenthesized term on the right-hand side is the positive saving at point in time 0. The second parenthesized term on the right-hand side is one plus the rate of interest that the savings will generate. Expanding the right-hand side shows that, for year 1, the decision maker, having saved, can consume an incremental amount equal to his or her current year's savings of $y_0 - c_0$, plus the interest that the current year's savings will earn, $(y_0 - c_0)$r.

On the other hand, by borrowing against next year's income, our decision maker can consume more than $y_0$ in the current year (at point in time 0). Here is how it looks for the decision to borrow:

$$c_0 - y_0 = \frac{y_1 - c_1}{1 + r} > 0 \tag{2.7}$$

The term on the left-hand side of equation (2.7) is the positive increment to year 0 consumption, and the term on the right-hand side is the present value (e.g., the discounted value of the future amount) of the amount $y_1 - c_1$ that will have to be paid back to the lender at time 1.

Let's recap. Our objective is to determine how much our representative household should optimally consume today (the current year) and in the future (next year). The answer depends on the household's wealth, on the relative price of the two consumption baskets, and on the decision maker's preference for current versus future consumption. The two income receipts, $y_0$ and $y_1$, establish the household's wealth position. But what is the price of consuming more now instead of a year from now, or of consuming more at time 1 instead of at time 0? The answer is implicit in equations (2.6) and (2.7).

Can you see it? If not, you will find out what it is in the next subsection.

### The Price of Current versus Future Consumption

Say that at time 0 you save one dollar and that your dollar earns interest for one year. Accordingly, at time 1 you will have one dollar plus the interest that your dollar has earned. You have saved and the market is paying you. For the one dollar that you have saved this year, you will have $1 + r$ dollars next year. That is the payment which you will receive. You will *be paid* $1 + r$ dollars next year for giving up one dollar this year.

What if at time 0 you have borrowed the present value of one dollar, which is $1/(1 + r)$. Because you have borrowed, you have to pay something back. The amount that you have to pay back at time 1 is $[1/(1 + r)]\,(1 + r) = 1$. This time you *pay* $1 + r$.

Because the borrowing rate equals the lending rate equals $r$, the price of shifting consumption between time 0 and time 1 is the constant, $1 + r$. Knowing this enables us to diagram the consumption opportunities available to our representative consumer. Exhibit 2.4 presents the diagram, with time 1 consumption ($C_1$) on the vertical axis and time 0 consumption ($C_0$) on the horizontal axis. The downward sloping line in the diagram shows that the decision maker can consume more at time 0 by consuming less at time 1 (and vice versa).

The downward sloping line is a straight line because its slope reflects the price of time 0 consumption relative to the price of time 1 consumption, and this price relative is the constant amount, $1 + r$. There are two ways to show that the slope equals $-(1 + r)$:

1. Point $A$ in Exhibit 2.4 shows that if 1 dollar is saved in the current period, consumption will be $(1 + r) * 1 = (1 + r)$ higher in the next period. Accordingly,

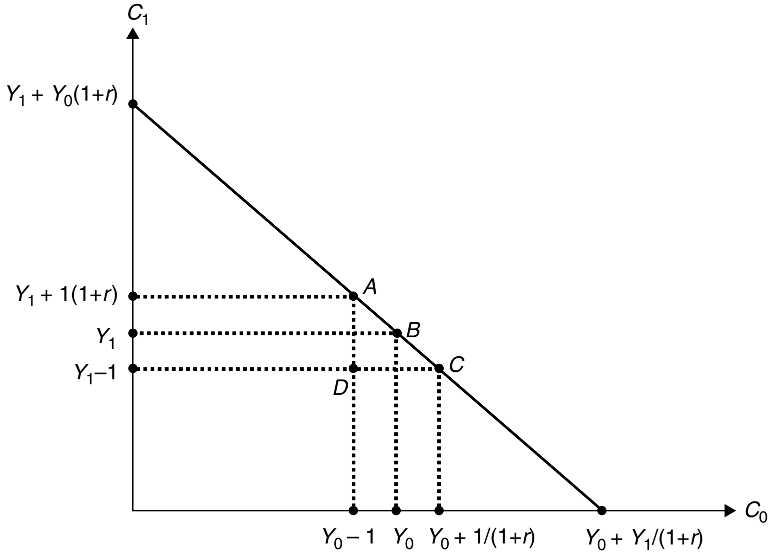


**EXHIBIT 2.4** An Individual's Budget Constraint for the Current Consumption, Future Consumption Decision

$$\frac{\Delta c_1}{\Delta c_0} = \frac{1+r}{-1}$$

Point C in the Exhibit shows that if consumption is reduced by 1 dollar next year, consumption will be higher by $1/(1 + r)$ in the current period. Accordingly, $\Delta c_1/\Delta c_0 = (-1)/[1/(1+r)] = (-1)(1+r)$, which is the same result.

2. The vertical intercept in Exhibit 2.4 shows the maximum amount that can be consumed at time 1 if consumption is zero at time 0: The amount is $c_1 = y_1 + y_0(1+r)$.

   The horizontal intercept shows the maximum amount that can be consumed at time 0 if consumption is zero at time 1: The amount is $c_0 = y_0 + y_1/(1+r)$.

   Dividing maximum $c_1$ by maximum $c_0$ shows that the slope of the negatively inclined straight line is $\Delta c_1/\Delta c_0 = -[y_1 + y_0(1+r)] \div [y_0 + y_1/(1+r)] = -(1+r)$, which is the same result.

Are you convinced? Do you agree that the slope of the line reflects the price of current consumption relative to future consumption? We hope so, because this is important. The slope of the straight line shows the rate at which current consumption can be exchanged for future consumption in the marketplace.

## The Budget Constraint

The straight line that we have just described is a *budget constraint*. It is a constraint because it demarks the alternative combinations of the two consumption bundles (in this case current consumption and future consumption) that are obtainable given the consumer's budget. All points in the positive quadrant that are on or below the constraint are obtainable, while no point above it can be reached. In the current case, the consumer's budget constraint is determined by his or her income receipts ($y_0$ and $y_1$) and by the price of borrowing or lending, which is $(1 + r)$. Specifically, $y_0$ and $y_1$ establish the location of the constraint that is set in Exhibit 2.4 by point *B*, and $1 + r$ sets the slope of the constraint.

Because the representative consumer gains utility from both current and future consumption (and given that the consumer's goal is to maximize utility), our happy consumer would never select a point that is not on the budget constraint itself. To assure yourself of this, arbitrarily pick a point such as *D* that is below the budget constraint as shown in Exhibit 2.4. In that exhibit, a vertical line from *D* to the constraint brings us to point *A*, while a horizontal line from *D* to the constraint brings us to point *C*. Any combination of the two consumption baskets anywhere on the constraint between *A* and *C* contains more of both future consumption and current consumption, and thus any point on the constraint between *A* and *C* must dominate point *D*.

## Indifference Curves

The budget constraint reflects two of the three key factors that determine consumer demand: wealth ($y_0$ together with $y_1$) and the price relative, $1 + r$. What remains to depict is the tastes of our representative consumer/decision maker.

For our current consumption, future consumption analysis, the utility function can be written simply as

$$U = f(c_0, c_1)$$

As with the *X*, *Y* trade-off, the utility function can be transformed into a family of indifference curves. Along each curve, utility is constant. The slope of a $c_0$, $c_1$ indifference curve at any point is

$$\frac{dc_1}{dc_0} = -\frac{MUc_0}{MUc_1} = -MRSc_0 \cdot c_1$$

As with an $X$, $Y$ indifference curve, a $c_0$, $c_1$ indifference curve is convex from below (i.e., moving along it from left to right, its negative slope becomes a smaller number in absolute value). The decreasing slope of an indifference curve is reasonable. It is biologically impossible for anyone to exist without consuming something in each of the two periods. Consequently, the decision maker's indifference curves can never intersect either axis. As $c_0$ is cut back toward zero, $c_1$ must rise toward infinity. Likewise, as $c_1$ is cut back toward zero, $c_0$ must go to infinity. In general, as our decision maker increasingly substitutes out of $c_1$ and into $c_0$ by successively giving up equal increments of $c_1$, he or she must receive ever-increasing increments of $c_0$ to keep his or her overall level of utility constant. Moving in the opposite direction, as our decision maker increasingly substitutes out of $c_0$ and into $c_1$ by successively giving up equal increments of $c_0$, he or she must receive ever-increasing increments of $c_1$ to keep his or her happiness meter constant. It would be a good idea for you to assure yourself that the convex indifference curves in Exhibit 2.2 satisfy this requirement.

## Interpretation

We all have different tastes for current consumption relative to future consumption. To some degree this is a matter of temperament. Some people like to save and abhor debt, while others like to spend *now* and are not troubled by debt. And, to a degree, one's attitude toward substituting current for future consumption (or vice versa) is life-cycle dependent. A young person might borrow to finance his or her education, people in the prime of their working life are likely to save for retirement and, after retiring, older people will live off their past savings (hopefully).

Nevertheless, despite this diversity, we can, rather amazingly, all have identical tastes for present relative to future consumption. That is, in a no-uncertainty, no-risk, frictionless world where everybody is maximizing his or her utility by consuming optimal amounts of the two consumption baskets, everybody is identical in one way—everybody is identical regarding the rate at which they are willing to swap one good for the other *on the margin*! How come, you might ask? Answer: because everybody's marginal rate of substitution of present for future consumption (the ratio of the marginal utilities) equals the price relative, $1 + r$, which is the same for all people. Thus, when the micro market is in equilibrium, everybody must have the same marginal rate of substitution and thus, on the margin, everyone's taste for present consumption relative to future consumption is the same. When this equality is achieved, there can be no further incentive to trade.

The result that everyone is identical on the margin shows the power that a micro market has to deliver a socially desirable outcome when the trading environment is frictionless, when risk and uncertainty do not exist, and when the decision makers are all price takers because no one individual participant is large enough in the marketplace to individually affect the interest rate, $r$.

Individuals who are adjusting their consumption patterns by borrowing or lending in the market are in effect trading with one another. People who on the margin differ in their tastes for present versus future consumption should trade with each other, and the trades that are made are mutually utility enhancing for the participants involved. Let's see how this works with the aid of an example.

Reciprocally stated, the trades that are utility enhancing are those that are made by participants whose utilities are different *on the margin*. On the other hand, equality on the margin is achieved via the trades that are made. To see this, consider our two consumers from Chapter 1, Alba and Harrison. What if neither borrows nor lends, but simply consumes income in the year when it is received? Assume that their marginal rates of substitution of present for future consumption if they do not trade are:

$$\begin{aligned}(MRS_{c_0 \cdot c_1})_{\text{Alba}} &= 1.08\\ (MRS_{c_0 \cdot c_1})_{\text{Harrison}} &= 1.06\end{aligned}$$

and that the price relative is

$$1 + r = 1.07$$

An $MRSc_0.c_1$ of 1.08 means that Alba's marginal utility of current consumption is 1.08 times her marginal utility of future consumption. Accordingly, she would be willing to pay up to \$1.08 at point in time 1 to borrow one marginal dollar at point in time 0 (but, of course, to do so, Alba would have to pay only the market rate of \$1.07).

Now onto Harrison. An $MRSc_0.c_1$ of 1.06 means that Harrison's marginal utility of current consumption is 1.06 times his marginal utility of future consumption. Accordingly, Harrison would be happy to lend one marginal dollar at any rate that gave him more than \$1.06 back at point in time 1. Accordingly, Alba could borrow a dollar from Harrison at any price higher than \$1.06 but less than \$1.08, and the arrangement would be mutually beneficial (they would both gain utility). Of course, with the market price set at \$1.07, the happy pair would each trade on the market at a price of \$1.07, and both exchanges would be utility enhancing.

In Exhibit 2.5 we show exactly how this exchange would be utility enhancing for each of them. To keep it simple, let Alba and Harrison both have the same income flows ($y_0$ and $y_1$). Their utility mappings,

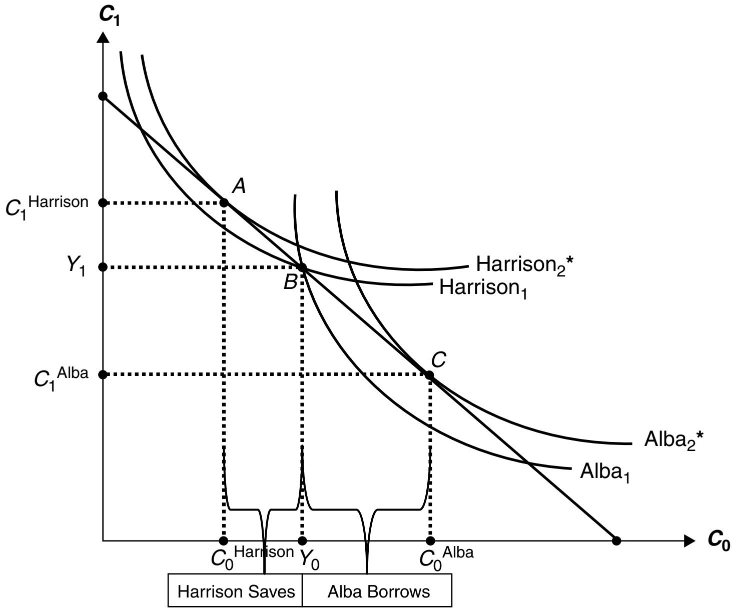


**EXHIBIT 2.5** Utility Enhancing Adjustments of Present Consumption and Future Consumption for Two Individuals

however, are different (because they have different tastes). The exhibit presents two indifference curves for each that are superimposed on the same graph with their names identifying which curves apply to whom. For each, the lower indifference curve is the curve that passes through the $y_0$, $y_1$ income point labeled *B*. Each curve is labeled by name (Alba or Harrison) and number (1 or 2); thus the curve labeled "Alba$_1$" is the (lower) indifference curve for Alba that passes through point *B*. The lower indifference curve for Harrison labeled "Harrison$_1$" also passes through point *B*. The Alba and Harrison indifference curves represent the utility that they each would achieve if neither of them saves, borrows, nor lends, but simply consumes each year's income in the year that it is received. But because neither of these indifference curves is tangent to the budget constraint at point *B*, both Alba and Harrison would benefit from altering their consumption patterns.

As we know from their respective marginal rates of substitution if they do not trade, and from the slope of the budget constraint, Alba's indifference curve is steeper than the budget constraint at point *B* and, moving from left to right, cuts it from above, while Harrison's indifference curve, being flatter, cuts the budget constraint at point *B* from below. Accordingly, Harrison will save and lend, and Alba will borrow. Alba will borrow just enough dollars for her *MRS* to fall to 1.07, and Harrison will lend just enough dollars for his *MRS* to rise to 1.07. These optimality conditions are satisfied when Harrison is at point *A* in Exhibit 2.5, and Alba is at point *C*.

The diagram has been drawn so that the amount that Alba borrows roughly equals the amount that Harrison lends. This is, of course, unnecessary; both of them should be thought of as trading impersonally in a marketplace where the aggregate demand for lendable funds of people like Alba intersects the aggregate supply of lendable funds from people like Harrison, and where the lending/borrowing rate of 7 percent is determined by the forces of demand and supply.

With Harrison relishing consumption bundle *A* and Alba feasting on bundle *C*, our two customers have identical tastes *on the margin* and no further, mutually beneficial trade would be possible. Under our assumed conditions, the micro market for borrowing and lending will have produced an excellent result indeed.

## THE CONSUMER CHOICE MODEL AND THE RISK-RETURN TRADE-OFF

We turn to another application of the household decision model in this section of the chapter. We consider the decision that an investor faces concerning the risk and return configuration of a financial portfolio comprised of risky stocks and a risk-free asset. This application is clearly important in its own right, and it further illustrates the more detailed thinking that is required to apply an abstract microeconomic model to a real-world trade-off situation. We start by considering what *risk* is all about.

*Risk* and its close associate *uncertainty* are pervasive in our lives and in the micro markets. Both are endemic to the equity markets. The two terms are similar in that they each imply that a given action can lead to more than one result, and you do not know exactly what will pan out. Aside from that, the two terms have different meanings. The term *risk* applies when the distribution of outcomes is identifiable and a probability number can be assigned to each of the possible outcomes. For instance, if you flip a coin, either of two results can happen: heads or tails. This property of a coin is known in advance and a probability can be assigned to each outcome. Assuming a fair coin, the probability of heads equals the probability of tails equals 0.5. Accordingly, betting on a coin flip involves risk, but in this case the term *uncertainty* does not apply.

Uncertainty means that you do not know all of the possible outcomes and their associated probabilities with precision. Uncertainty can typically be diminished though not necessarily eliminated by gathering more information. Risk cannot be equivalently reduced by acquiring more information per se, but it can be controlled by proper diversification, portfolio balancing, and otherwise hedging risk positions (e.g., with financial derivatives).

Both risk and uncertainty have been analyzed and written about extensively in the economics literature. In this section of the chapter, we deal with risk, but not with uncertainty. To do so, we focus on one specific and very important issue: how to resolve the risk, return trade-off that characterizes alternative investments in assets which offer future returns that are probabilistic. We will consider an equity portfolio. The consumer choice model developed in the previous two sections will be used to show how an investor in this micro market can select a portfolio that offers an optimal combination of risk and expected return given that person's taste for risk. As in the previous section, we will continue to make the simplifying assumption that the investor is operating in a frictionless market environment. We trust that you will see how this abstraction from reality enables us to unearth some major, underlying forces that are at work in the micro markets for financial capital.

### The Meaning of Risk Aversion

How do people feel about risk? How do you feel about risk? For the most part, most people do not like to take chances if they do not have to, or if they are not adequately compensated for doing so. There are exceptions, of course. Some people like to bet on games, contests, and whatever else. Some people go to Monte Carlo and Las Vegas, Nevada. Some want to take in the scene and see all the action. Some like the fun of the game. Some people might even expect to win. This risk-seeking has the flavor of a consumption good. And, in some cases, expectations of winning may not be founded on totally objective reality. None of this has anything to do with a microeconomist's analysis of risk-averse behavior.

Let's get real. Our focus on risk aversion is in the context of rational decision making about investments that, by their future payouts, will add to or subtract from your wealth (and therefore increase or reduce your future consumption possibilities). We assume that your concern is with your wealth position, not with any attitudes, either positive or negative, about the excitement of risk-taking itself. Your attitude about risk has to do only with the financial payouts that you might receive (or lose).

Your taste for risk can be defined with respect to your willingness to take a fair bet. Here is an example of a fair bet: For $1,000, you can buy the right to win \$2,000 with a 0.5 probability, and you walk away with nothing if you do not win. Are you indifferent about the bet? Will you take it? Or will you simply turn your back on the bet and walk away? Think about it. In the meantime, here is how a microeconomist would characterize your answer: If you are indifferent about the bet, you are *risk neutral*; if you would take the bet, you are a *risk seeker*; and if you would walk away, you are a *risk averter*.

We will take a guess: Most people are risk averse, and we assume that you are, too. This does not mean that you will not accept risk. What it means is that, as a risk averter, you will accept risk only if you are rewarded for doing so. The reward is the payment that you can get in the marketplace for accepting risk. The more risk averse you are, the higher the payment that you would require to take any given risk.

Here is another example of a risky situation, one that has nothing to do with the return on a financial investment. Assume that you are driving west on a three-lane (in each direction) highway that you have not traveled before. It is late afternoon and the sun, along with fairly heavy traffic, is making your life a good deal less pleasant. You see a sign stating that the exit to another highway that you will need to take is one mile ahead, but you do not know if the exit is to the right or the left. Assume that the probabilities are 50, 50. You are cruising along at 60 miles per hour, and with one minute to go, you can easily change your lane. With 30 seconds or less to go, lane changing becomes increasingly challenging. The question is, which lane should you position yourself in well before you see the exit sign: right, left, or center?

If you choose the right lane or the left lane, you have a 50 percent chance of not having to change your lane to exit, and a 50 percent chance of having to change two lanes. The expected number of lane changes is .50 (2 lanes) + .50 (0 lanes) = 1 lane. If you choose the middle lane, you will with certainty have to change one lane. So, the expected value of the risky selection equals the certainty value of the riskless selection. What will your decision be? The sun is in your eyes, the traffic is heavy, and having to change two lanes quickly is definitely more dicey than having to change one. If this is how you see it, you should position yourself in the center lane. Doing so is risk-averse behavior. Recognizing this, we can ask what the probability must be that the exit is to the right (or left) for you to optimally choose the right- (or left-) hand lane. Does it make sense to you that the difference between this higher probability and .50 would be compensation to you for taking a risk?

The reason why most people would walk away from a betting situation such as the two we have just described is that the payout is structured so that you are paid nothing for taking the bet. Let's return to the financial bet and look at it this way. By walking away, you can keep your \$1,000 with certainty; by taking the bet your expected winnings are $0.5(0) + 0.5(\$2{,}000) = \$1{,}000$ and you have to pay \$1,000 to play, so your expected gain is zero. Accordingly, the expected payout from taking the bet just equals the amount that you can have with certainty. And, all else equal (*ceteris paribus*), as a risk averter, you prefer certainty. Accordingly, if you are risk averse, you will keep your \$1,000 and walk away.

In terms of the dollars that can be won or lost, we have just described a *fair bet* (the lane selection example also describes a fair bet). With a fair bet, the expected dollar return from the bet, all cash flows considered, is 0. But why will somebody purposefully walk away from a fair bet? If you *on expectation* will break even, why not just be neutral about the whole thing? The answer to this question will explain what risk aversion is all about, what it really means about a person's tastes.

For most of us, \$1,000 is a lot of money. How would you feel if you were to win \$1,000? Very good, no doubt. But contrast this feeling with what your attitude would be if you were to lose \$1,000. How would your *incremental happiness* if you win stack up against your *incremental misery* if you lose? This question brings us back to an equivalent situation that we considered in Chapter 1 while discussing marginal analysis.

Back in Chapter 1, our friend Alba was thinking of her \$100,000 portfolio and considering how she would feel if her wealth, after the passage of one year, was valued at \$90,000 (negative growth), or at \$100,000 (zero growth), or at \$110,000 (good growth), or at \$120,000 (strong growth). We described her emotions as, respectively, total disgust, disappointment, satisfaction, and total delight. Then we went further. We suggested that a utility value could be assigned for each outcome from the lowest (\$90,000) to the highest (\$120,000), where the higher outcomes got higher utility values. We then suggested that the differential utility achieved by being at \$100,000 instead of \$90,000 would be greater than the differential utility achieved by being at \$120,000 instead of \$110,000. Likewise, in the current betting case, we suggest that the utility lost if you were to lose the bet and be down \$1,000 is greater than the utility gained if you were to win the bet and be ahead by \$1,000. Would that depiction of the situation hold for you? If so, you are, indeed, a risk averter.

To assess a bet that in terms of its dollar payouts is a fair bet, we translate the dollar amounts into utility values. A most important thing then happens. With diminishing marginal utility for wealth, the bet, in utility terms, is no longer a fair bet. When assessing the payouts in utility terms, you expect to lose. Moreover, the more rapidly your marginal utility of wealth decreases as your wealth rises, the more you expect to lose. That is why the bet in dollar terms must have a positive expected value for you to take it. The positive dollar expectation is your compensation for taking the risk. The compensation is called a *risk premium*.

We will discuss the risk premium in greater detail shortly. At this point, suffice it to say that, the more risk averse you are, the higher this risk premium must be. And here is a punch line: Your attitude toward risk has nothing to do with the process of betting per se; rather, it is a manifestation of your tastes, on the margin, for wealth (or income).

Let's quickly circle back to the lane selection example. We saw a few paragraphs ago that, when approaching a highway exit which with equal probability can be from either the right lane or the left lane, the optimal position for the risk-averse driver to be in is the center lane. Is this solution, like the dollar-bet example, being driven (no pun intended) by diminishing marginal utility? No, not this one. This one is a case of increasing marginal costs. Do you see it?

Let's return again to monetary risks. In Exhibit 2.6 we illustrate three utility functions (of wealth). The higher one is curving upward at an increasing rate (the marginal utility of wealth is rising with wealth); this is the utility function of a risk seeker. The middle utility function is linear (the marginal utility of wealth is constant); this is the utility function of someone who is risk neutral. The lowest utility function is rising upward at a decreasing rate

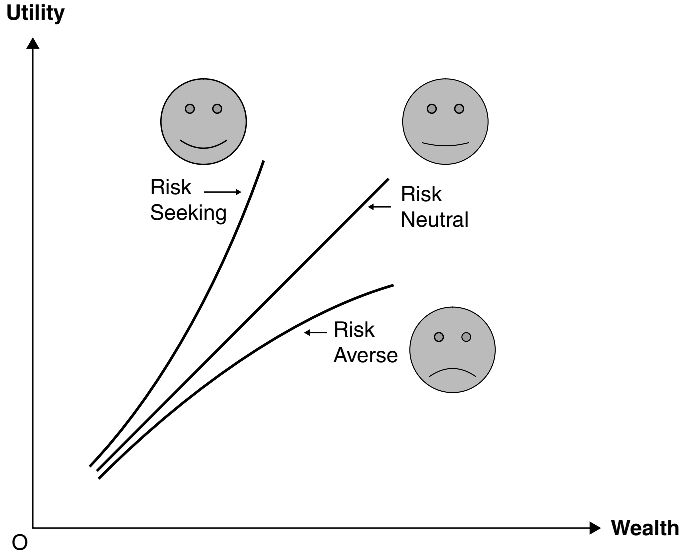


**EXHIBIT 2.6** Utility of Wealth Curves for a Risk-Seeking, a Risk-Neutral, and a Risk-Averse Individual

(the marginal utility of wealth is falling as wealth increases); this is the utility function of a risk averter. When it comes to making wise decisions about how to put your savings to work today so that you might consume more in the future, we will assume that your utility from future wealth (and therefore from future consumption) is increasing, but at a decreasing rate. That is, we will assume that you are a risk averter. This probably is a good assumption. When it comes to sensible investing, most people are.

### Returns and Expected Returns

*Returns*, *expected returns*, and *risk* are fuzzy terms until we give them precise definitions. We deal with returns and expected returns in this subsection, and turn to risk in the next. After that, we proceed to another closely related term: a *risk premium*. As we discuss these terms, it is helpful to keep the following four questions in mind:

1. When we refer to the risk of holding a financial asset, what exactly do we mean by "risk"? How should risk be measured?
2. We will show how risk is measured with reference to an asset's returns, not its price, per se. So, how do we measure returns?
3. What is the price of risk? Alternatively stated, how does the micro market for risky assets compensate one for investing in these assets and thereby accepting risk?
4. What exactly is a risk premium and how does it relate to the price of risk?

Let's start at the beginning, with the measurement of returns. A stock's value at any specific point in time is its price at that point in time. The return for a stock is its *price change* (typically expressed as a percent) over a period of time. For instance, if stock A has been bought at $10 a share and its price rises to $11 one year later, the realized annual rate of return for that stock is 10 percent. In the equity markets, returns typically are further enhanced by the payment of *cash dividends* and, to measure returns properly, share prices must also be adjusted for *stock dividends* and *stock splits*. For simplicity, let's assume that these factors are absent, that the sole contributor to returns is the change in the stock's price over time. Then, one possible measure of return for any span of time $T - 1$ to $T$ is a stock's percentage price change,

$$\frac{(P_T - P_{T-1})}{P_{T-1}}$$

The lowest value that a return so measured can attain is $-100$ percent because, with limited liability, you cannot lose more than your total investment (i.e., share prices cannot fall below zero, and you have no other liabilities with respect to owning them). This means that the return measured as a percentage price change cannot be normally distributed (the standard bell-shaped distribution is not bounded: It extends all the way from negative infinity to positive infinity).

This little complexity is easily dealt with. Adding one to the percentage price change and taking logs converts the arithmetic return into a *logarithmic return*. In the limit, as $P_T$ approaches zero, the log of 1 plus the percentage price change approaches minus infinity, which means that the log return is not bounded from below. Empirical studies have shown that log returns are pretty much normally distributed. As you will appreciate in the next subsection, a normal distribution has very desirable properties when it comes to the analysis of risk. So, for empirical work, a return from point in time $T - 1$ to point in time $T$ should be defined as

$$r_T = \ln\left[1 + \frac{(P_T - P_{T-1})}{P_{T-1}}\right] \tag{2.8}$$

where ln denotes a natural logarithm and $r_T$ is a log return.

For our purposes, we will simply refer to $r_T$ returns, and will assume them to be normally distributed. Unless you are undertaking an empirical analysis, it is not necessary to keep referring to $r_T$ as a logarithmic return.

For decision making, we formulate expectations concerning the future. Our expectations, however, are usually based on observations from the past. What can we learn from data that have been generated in the past? To the extent that returns distributions are reasonably stationary over time (i.e., do not keep shifting all around), a lot can be learned. Say we have a one-year investment horizon. Visualize our annual investment returns as being randomly drawn from a distribution of possible one-year returns, and assume that the distribution has indeed been stationary. In other words, let the alternative outcomes and their associated probabilities be known, just as they are for a coin flip. Let the set of possible outcomes be described by a mean and a variance.

A normal distribution is described by just two parameters, its *mean* and its *variance*. The mean establishes the location of the distribution, and the variance describes the dispersion of outcomes around the mean. Think of it this way: The mean return is what the market will pay you, *on expectation*, for accepting risk, and the variance of returns is a measure of that risk. If the normal distribution (note, for simplicity we said "normal," not "lognormal") does not shift around over time, the return that we *expect to receive* from the next draw from the distribution (e.g., the next annual return) is the mean of the distribution of past annual returns.

As is standard in statistics, the mean (and the variance as well) can be estimated from draws (that is, from actual returns) that were realized in the past. Thus we can express the mean as

$$\mu_r = \mathrm{E}(r) = \frac{1}{N}[r_1 + \cdots + r_N] \tag{2.9}$$

where $\mu_r$ [and $\mathrm{E}(r)$] is the expected return, and $r_1 + \cdots + r_N$ is the sum of $N$ observations from the first observation, $r_1$, through the last, $r_N$.

Annual realized returns, however, need not be used to estimate an average annual return. This is fortunate, for it would take us too far back into the past to collect enough data to have a sufficiently large sample on which to base our estimates (some companies might not even be old enough to provide annual data that go back far enough). Rather, one year's worth of data might be enough. We could, for instance, compute the average weekly return and scale it up to an annual expected return by multiplying by 52.

## The Measurement of Risk

Risk is inherent in the volatility of stock prices that we see not only over relatively long periods of time in the equity markets, but also on an intraday basis. We talk about *price volatility*, but measure volatility by the variance of returns (not by the variance of prices). Normal returns distributions that are stretched out more widely have larger variances, while normal returns that are distributed more tightly around their means have smaller variances. A variance is defined as the *expected squared difference* between a randomly selected observation and its mean. Using data from the past, we can measure returns variance as

$$Var(r) = \mathrm{E}(r - \mu_r)^2 = \frac{1}{N}\left[(r_1 - \mu_r)^2 + \cdots + (r_N - \mu_r)^2\right] \tag{2.10}$$

Like the mean, the variance of a longer period return (an annual return, for instance) can be estimated from shorter period returns (a weekly return, for instance). That is, you can obtain an estimate of the weekly variance of log returns and annualize it by multiplying by 52. You have to be careful with this one, however, because this exact relationship would be distorted by any correlation patterns that might be present in weekly returns data (the expected return would not be affected by any correlation pattern).[3]

---

[3] The reason for this complexity regarding the variance term is explained by the following technicality. While the sum of the two terms, $A$ and $B$, is $(A + B)$, the variance of the two terms combined is $Var(A) + Var(B) + 2^{*}Cov(A, B)$, where the covariance term reflects the correlation between $A$ and $B$.

A *standard deviation*, which is the square root of variance, is another measure of dispersion. Standard deviations and variances can be used fairly interchangeably because they are essentially different ways of looking at dispersion (much as the length of a table can be measured in either feet or inches). One useful property of a standard deviation is that it is expressed in the same units of measurement as the underlying data, whereas variance, the squared standard deviation, is expressed in terms of the units squared. For instance, if the standard deviation of annual returns for a particular stock is 7 percent, the corresponding variance is 7 percent squared or 0.0049. Another important property is that the standard deviation can be understood in probability terms. For instance, 68 percent of the draws from a normal distribution lie within one standard deviation of its mean, 95 percent of the draws are within two standard deviations of the mean, and so forth. There is one more thing: some mathematical applications are simpler when the standard deviation is used rather than the variance. As you will see, we use the standard deviation to analyze the risk, return trade-off.

## A Risk Premium

As we have said, a risk averter will accept risk only if paid to do so. The payment received is *a risk premium*. This can be understood in relation to the coin-flip example. We have already established the following. The probability of a fair coin landing heads up is 0.5. Suppose that you put down $1,000 for the chance of being paid $2,000 if the coin lands heads but nothing if it lands tails. Your expected winning from the risky bet is $1,000, which just equals the cost of making the bet. Because the expected winning equals the $1,000 cost (which is known with certainty) of placing the bet, we have described a fair bet, the risk premium is zero, and a risk averter will not take the bet.

But what if you are paid $2,100 if the coin lands heads up? In this case, your expected return is $0.5 \times \$2{,}100 = \$1{,}050$, your dollar risk premium is $50, and your percentage risk premium is 5 percent. Is that payment enough for you to take the bet? The answer depends on how risk averse you are. If you are not very risk averse, a 3 percent risk premium might be more than enough; but if your risk tolerance is relatively low, you might take the bet only if the percentage risk premium is 8 percent or more. It is your call.

*Your* risk premium describes *your risk aversion*. Your risk premium is the added return that *you require* to take a bet and not walk away. A *market* risk premium is a *compensation that is offered* in the marketplace. For instance, if the expected return on a market portfolio is $r_M$ and the interest rate on a riskless, short-term government bill is $r_f$, the risk premium that the market is paying is $E(r_M) - r_f > 0$. Let's say this in another way.

The riskless return, $r_f$, is the return that the market pays you to compensate you for postponing consumption, and $E(r_M) - r_f > 0$ is the *additional return* that you get for investing in the *risky* asset. That is, $E(r_M) - r_f > 0$ is the return that the market pays you for accepting risk (i.e., it is the risk premium). Got it?

The risk premium offered by the market is like any other price. It is set on the market by the forces of demand and supply. We continue to assume that no investor is big enough to affect the market price of risk, that we are all price takers with respect to the market's risk premium, $E(r_M) - r_f$. On the other hand, an individual investor's marginal disutility of risk will increase as he or she accepts more risk. In other words, if an investor were to accept little risk, it is likely that his or her marginal disutility of risk (as measured by the premium required for a marginal increase of risk) is less than the constant risk premium offered by the market. However, if the investor were to accept a lot of risk, his or her marginal disutility of risk would increase and, eventually, would be greater than the market's constant risk premium.

Returning to the coin example, let's say that the representative investor requires a positive risk premium of 250 dollars. This can be obtained if \$2,500 rather than \$2,000 is the payout for winning the bet, for then the expected payout is $.5(0) + .5(2{,}500) = 1{,}250$, and the risk premium in dollars (the expected payoff if the bet is taken less what you have paid to play the game) is $1{,}250 - 1{,}000 = 250$. In percentage terms, the calculation works out as follows. The bet offers you a 50 percent chance of losing 100 percent and a 50 percent chance of winning 150 percent. So, we can write $0.5(-100 \text{ percent}) + 0.5(+150 \text{ percent}) = 25$ percent, where 150 percent is the net return that you make if you win the bet, given that you have paid \$1,000 to play. To wrap this up, the expected return of 25 percent that you require for playing the game is your risk premium because the expected return for the fair bet was 0.0 percent.

Back to the stock market: Suppose that you are considering the number of shares to buy of a broadly diversified market portfolio (call it, if you wish, an index fund), and that the investment offers a risk premium of $E(r_M) - r_f$. The more shares you buy, the more risk you are acquiring and, for you to accept a greater risk, the risk premium paid by the market would have to be larger. Recognizing this, we can differentiate between your

**Total risk premium** (in percentage terms): what you have to be paid to hold a given number of shares.

**Average risk premium** (in percentage terms): your per share risk premium, which is your total risk premium divided by the number of shares that you are holding.

**Marginal risk premium** (in percentage terms): the amount by which your total risk premium goes up per each incremental increase in the number of shares that you hold.

As the risk acquired by a risk averter increases, the total risk premium increases at an increasing rate, which means that the average risk premium also goes up, and the marginal risk premium goes up even faster.

The increasing marginal risk premium is of particular importance for our analysis of this micro market. We are jumping ahead of our story a bit, but here is a punch line: If the representative investor's share holding of the risky market portfolio is too small, his or her required marginal risk premium is less than the risk premium the market is paying, and more shares should be held. On the other hand, if the investor's share holdings are too large, his or her required marginal risk premium is greater than the market's risk premium, and fewer shares should be held. But when the representative investor is holding just the optimal number of shares, the decision maker's marginal risk premium exactly equals the market's risk premium, $E(r_M) - r_f > 0$, and the investor should have no further desire to trade.

## The Risk-Return Trade-Off

We considered the trade-offs between $X$ and $Y$, and between current consumption and future consumption, in the previous two sections of this chapter. By saving out of current income ($y_0$) and acquiring a financial asset, our representative consumer is able to increase his or her future consumption. Because more than one financial asset will generally be held by the investor (given the rigorous foundation established by modern portfolio theory, we know that it pays to diversify), let's think in terms of a financial *portfolio*.

To understand the risk-return trade-off in a relatively simple context, we can consider portfolio choice as it applies to the determination of an optimal balance between just two financial assets, much as we previously separated the consumption choice into just two baskets, first $X$ and $Y$, and then consumption today and consumption tomorrow.

The two financial assets that we will work with are:

1. A risky market portfolio that you might think of as a well-diversified mutual fund.
2. A riskless asset, such as a short-term government bill (e.g., Treasury bills).

By "risk-return trade-off," we literally mean risk and *expected return*. Our first objective is to obtain the menu of efficient risk and expected return combinations that can be realized by changing the proportion of our holdings that is in the risky market portfolio, and the proportion that is in the risk-free asset. The representative investor will select an optimal risk, expected return combination from this menu. The menu will be represented graphically by an upward sloping straight line that plays a role similar to that played by the budget constraint in our previous analyses of the $X$, $Y$ trade-off and the present consumption, future consumption trade-off. You will see that the pitch of the upward sloping line reflects the price of risk. We refer to the upward sloping line as the *capital market line*.

We use the following cast of characters to obtain the capital market line:

$\mathrm{E}(r_m)$ = the expected return on the risky market portfolio.

$\sigma\,(r_m)$ = the standard deviation of returns on the risky market portfolio.

$r_f$ = the known return on the risk-free asset.

$\mathrm{E}(r)$ = the expected return on the portfolio that contains both the risky market portfolio and the risk-free asset.

$\sigma$ = the standard deviation of returns on the portfolio that contains both the risky market portfolio and the risk-free asset.

$w_m$ = the proportion of the individual's total financial holdings that is invested in the risky market portfolio. Specifically, $w_m = P_m N_m / W$, where $P_m$ is the price of one share of the risky market portfolio, $N_m$ is the number of shares held, and $W$ is the investor's wealth.

$w_f$ = the proportion of the individual's financial holding that is invested in the riskless asset. Note that the two weights must add to one and thus $w_f = 1 - w_m$.

The capital market line shows the relationship between the expected return on the portfolio that contains both the risky market portfolio and the risk-free asset [this is $\mathrm{E}(r)$] which is on the vertical axis), and the standard deviation of returns for that combined portfolio (this is $\sigma$, which is on the horizontal axis).

The expected return on the combined portfolio is

$$\mathrm{E}(r) = w_f r_f + w_m \mathrm{E}(r_m) = (1 - w_m) r_f + w_m \mathrm{E}(r_m) \qquad (2.11)$$

Because the risk-free asset is riskless, its standard deviation is zero. Thus the standard deviation of the combined portfolio is[4]

---

[4] To see this, write $\sigma^2 = w_m^2\sigma_m^2 + w_f^2\sigma_f^2 + 2w_m w_f \sigma_m \sigma_f \rho_{m,f}$. Recognizing that, for the risk-free asset, $\sigma_f = 0$, we get equation (2.12).

$$\sigma = w_m \sigma_m \tag{2.12}$$

From equation (2.12) we have

$$w_m = \frac{\sigma}{\sigma_m} \tag{2.13}$$

Substituting into equation (2.11) and rearranging gives us the equation for the capital market line,

$$\mathrm{E}(r) = r_f + \left[\frac{\mathrm{E}(r_m) - r_f}{\sigma_m}\right]\sigma \tag{2.14}$$

This equation is presented graphically in Exhibit 2.7.

Equation 2.14, which shows E($r$) as an increasing function of $\sigma$ [note that $r_f$, E($r_m$), and $\sigma_m$ are parameters], has an interesting interpretation. The left-hand side tells us what the expected return should be for any level of risk as measured by $\sigma$, the standard deviation of the combined portfolio. The right-hand side shows that this expected return can be decomposed into two components:

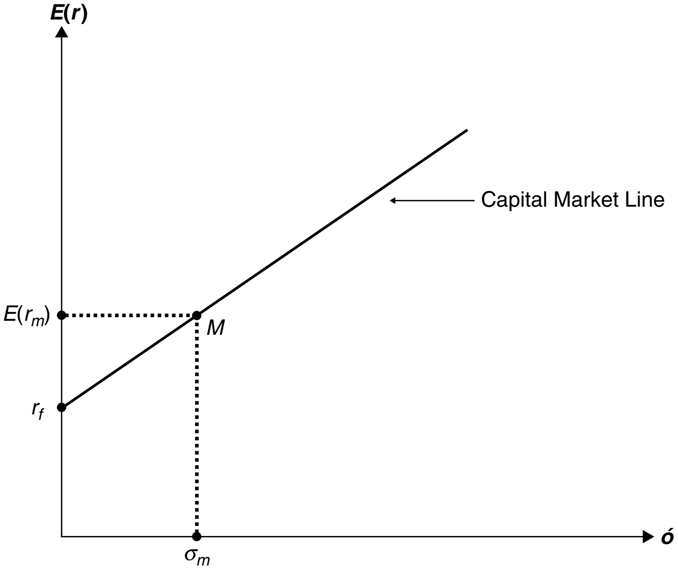


**EXHIBIT 2.7** The Capital Market Line

1. The compensation for postponing consumption, $r_f$, (we focused on this decision in the preceding section)
2. $\left[\left(E(r_m) - r_f\right)/\sigma_m\right]\sigma$

The second component is a risk premium per unit of risk in the risky market portfolio, times $\sigma$, the amount of risk in the combined portfolio that the representative investor has chosen and thus must be compensated for.

Let's recap. Bringing it all together, the investor selects a risk-return combination, $\sigma$ and E($r$). What does E($r$) compensate for? Answer: the cost of postponing consumption (the $r_f$ component does this) plus the cost of accepting the risk represented by $\sigma$ (the $\left[\left(\mathrm{E}(r_m) - r_f\right)/\sigma_m\right]\sigma$ component does this).

How does the investor change his or her risk-return combination? By moving up or down along the capital market line, a movement that is accomplished by changing the portfolio weights. Plugging into equation 2.12, we see that:

1. If $w_m = 0$, then $\sigma = 0$ and E($r$) equals the risk-free rate.
2. If $w_m = 1$, then $\sigma = \sigma_m$ and $\mathrm{E}(r) = \mathrm{E}(r_m)$.
3. If $w_m = 0.5$, then $\sigma = ½\, \sigma_m$ and $\mathrm{E}(r) = r_f + ½[\mathrm{E}(r_m) - r_f)]$

and so on.

The capital market line in Exhibit 2.7 has an intercept of $\sigma = 0$, $\mathrm{E}(r) = r_f$, and it goes through the point $\sigma = \sigma_m$ and $\mathrm{E}(r) = \mathrm{E}(r_m)$, which we have labeled M. Then it keeps on going. What does this mean? For any point along the line from the intercept to point M, the investor puts part of his or her funds in the risk-free asset, and part in the risky market portfolio. In this range, the investor is *lending* at the risk-free rate. For any point on the line past M, the investor's portfolio weight in the risky market portfolio is greater than unity. What is going on? Along the line beyond M, the investor is *borrowing* at the risk free-rate and leveraging up his or her expected return on the risky market portfolio. That is what a weight greater than unity implies. At the same time, the weight in the risk-free asset is negative (the investor has borrowed, not lent), and the weights still sum to one.

The capital market line (equation 2.14 and Exhibit 2.7) can be thought of as a budget constraint in that, given a market price (in this case, the price of risk), it separates obtainable from unobtainable points. A technical difference is that the wealth being allocated by the investor is not explicit. And neither need it be. The investment decision is formulated in terms of percentage allocations between the two assets, one riskless and the other risky. Once the percentages have been determined, they can be applied to the total amount of funds that the decision maker has decided to invest in these two financial assets. The available funds could be related to the amount of income saved at time 0, which is $y_0 - c_0$, as discussed in the preceding section (although we did not consider risk in that earlier discussion).

Another difference between the capital market line and a typical budget constraint is that the capital market line is upward sloping while the standard budget constraint is downward sloping. As we have seen in the previous two sections, budget constraints and indifference curves are typically applied in consumer choice models to analyze allocations between items (e.g., $X$ versus $Y$, or current consumption versus future consumption) that the consumer gets positive utility from and thus wishes to consume in the largest amounts possible. Alternatively stated, the typical analysis is of a *good-versus-good* trade-off. In contrast, our current application considers a *good* (expected return) versus *bad* (risk) trade-off.

Thus far, we have not explained how the risky market portfolio and the capital market line may themselves be identified, a task that we next attend to. In so doing, we will now locate the budget constraint. Interestingly, given the frictionless environment that we are operating in, it will turn out that all investors face the same budget constraint.[5]

To locate the risky market portfolio and budget constraint, first take a look at Exhibit 2.8. The exhibit shows three candidates for the risky market portfolio, along with an arrow pointing in the northwest direction. The arrow helps to keep us oriented regarding the direction (it is to the northwest) in which the investor would ideally like to move to gain more utility. Why northwest? Because returns are good, but risk is bad, because the investor would like to get more expected return (move north) and, as much as possible, to reduce risk (move west).

In markets, choice exists regarding the selection of a risky market portfolio. We can simplify the analysis of this choice by introducing the notion of a risky market portfolio that every individual stock will be represented in. Suppose that three mutually exclusive candidates for the risky market portfolio exist (i.e., each individual investor can put part of his or her funds in the risk-free asset and part in one of the three candidate portfolios). Question: Will one of the three candidate portfolios be preferred over the other two by all of the investors, irrespective of their individual tastes for risk? The answer is "yes"—let's see why.

---

[5] The formulation also requires that investors have *homogeneous expectations* (i.e., they must all agree on the risk and return parameters of different stocks and portfolios). If different investors do not form identical expectations, the term we use is *divergent expectations*. We discuss homogeneous versus divergent expectations in greater detail in Chapter 7.

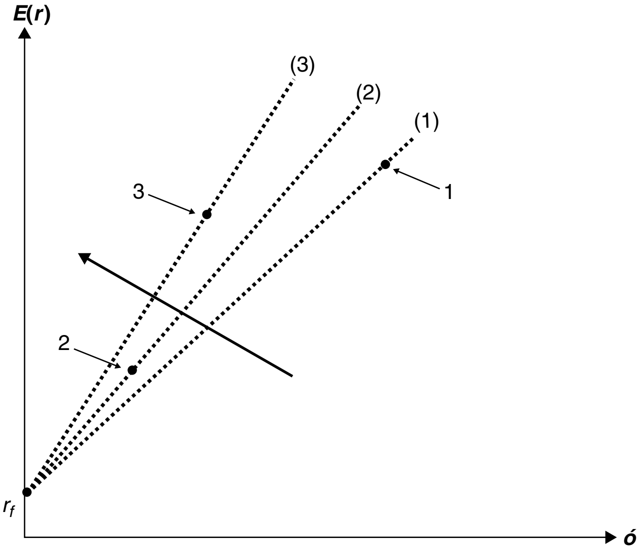


**EXHIBIT 2.8** Selection of the Risky Market Portfolio That Maximizes the Slope of the Capital Market Line

As we have noted, Exhibit 2.8 displays three candidate portfolios with risk and return parameters that locate them at points 1, 2, and 3, respectively. Note that all three would fall on an upward rising curve and thus, without further structure, it is not clear which would be selected. Which one would you choose? To answer this, look at the three straight lines that have been drawn in Exhibit 2.8 from the $E(r)$ intercept, $r_f$, through each of the three points, respectively. Portfolio 3 lies on the steepest of the three lines, the line labeled 3. Line 3 shows the highest $E(r)$ for any value of $\sigma$, and the lowest $\sigma$ for any value of $E(r)$. Accordingly, you can get further to the northwest by selecting portfolio 3 (and, in so doing, by selecting line 3) over the other two portfolios to combine with the risk-free asset. And so, for you and every other investor, regardless of any of your tastes for risk, risky portfolio 3 dominates risky portfolios 1 and 2. Risky portfolio 3, therefore, is the winner: It is the risky market portfolio. This risky market portfolio, along with the risk-free rate of return, locates the budget constraint. The constraint, it turns out, is the capital market line.

Just how the various stocks' individual weights are set and what this implies about how the individual shares are priced, is a topic addressed in finance courses that deal with a formulation called the *Capital Asset Pricing Model* (or CAPM, pronounced "CAP-M" by those who are familiar with the model). CAPM is based on the assumption that investors, as a group, have homogeneous expectations, that markets are frictionless, and that, in

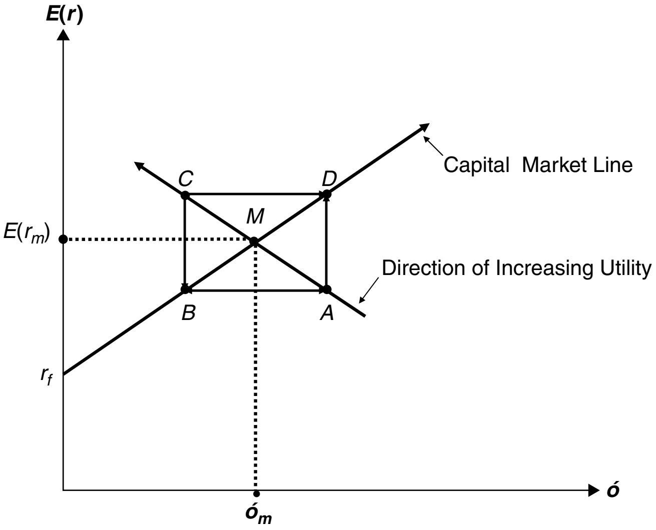


**EXHIBIT 2.9** The Capital Market Line, an Inferior Risk and Return Combination (*A*), and an Unattainable Risk and Return Combination (*C*)

the frictionless environment, every investor holds shares of every single one of the traded stocks.[6] Is it reasonable to assume that investors will all hold each and every risky financial asset in the single portfolio labeled *M* in Exhibit 2.7? In the real world, no. In our theoretical model, of course it is. Remember, we are operating in a frictionless environment.

Now that you know how the risky market portfolio labeled *M* in Exhibit 2.7 is established, we will continue to deal with the risky market portfolio only, and stop short for now with going further into the CAPM. Exhibit 2.9 replicates Exhibit 2.7 with four additional points added. In keeping with the arrow that is pointing to the northwest, all points below the capital market line (e.g., point *A* in Exhibit 2.9 and points 1 and 2 in Exhibit 2.8) are inferior. Further, no point above the line (e.g., point *C* in Exhibit 2.9) is attainable.

---

[6] The CAPM has a further feature. It can be shown in greater detail that the risky market portfolio that contains all stocks in proportions that maximize the expected return per unit of risk dominates all other single stock combinations and portfolio combinations for all investors. CAPM, therefore, delivers a market-wide equilibrium solution for individual investors' asset holdings and market prices. As such, it satisfies the requirement that, when the broad market is in equilibrium, all outstanding shares for all publicly traded securities be held in the investors' portfolios. Note that, if any stock was not fully included in the risky market portfolio, its price would fall because of insufficient demand, and its expected return would correspondingly rise until that stock was also purchased and held in the appropriate proportion in the market portfolio. Thus, after the dust has settled, no shares will be left unheld. All markets will clear and an equilibrium structure of asset prices will have been attained.

Point *A*, an arbitrarily selected point below the line, is clearly inferior to *B*, which is on the Capital Market line (*B* offers the same expected return with less risk). *A* is also inferior to *D* which is on the Capital Market line as well (*D* offers a higher expected return for the same risk). Compared to *A*, all points on the line between *B* and *D* offer more expected return and less risk, and thus are preferred. Point *C* (which has been selected with reference to points *B* and *D*) is not obtainable: *B* shows the *maximum* expected return that the market will offer for the level of risk at point *C*, and *D* shows the *minimum* risk that can be achieved for the higher expected return offered by point *C*. The maximum and the minimum are constraints that make *C* unattainable.

Put it all together and this is what we have for this frictionless micro market: The investor's optimal risk, return combination must be a point on the capital market line (which we have seen is the budget constraint). Question: Which point will the representative investor select? Answer: the optimal point, of course. The optimum point is set by the highest indifference curve the investor can reach.

### Risk and Return Indifference Curves

In discussing attitudes toward risk, we have already established that investors are generally risk averse. In Exhibit 2.6, we have previously displayed a representative utility of wealth function for the risk-averse investor. The key characteristic of this function is that it is rising at a decreasing rate (i.e., it is concave from below). Because of this, the investor has to be compensated for accepting risk. Furthermore, the more the investor is compensated, the more risk he or she will accept. All of this translates into risk, return indifference curves that are upward sloping and convex from below. Three such curves are displayed in Exhibit 2.10. They represent a partial mapping of the investor's risk and return utility mountain.

The *positive slope of each curve* reflects the fact that while expected return is a good, risk is a bad. The *convexity of the curves* reflects the fact that the investor's disutility from accepting risk increases on the margin as his or her portfolio acquires more risk. To understand the convexity more clearly, pick one of the three indifference curves, scan it from left to right, and note that, for utility to remain constant as $\sigma$ is increased by equal amounts, E($r$) has to be increased in increasing amounts because the investor's marginal distaste for risk increases.

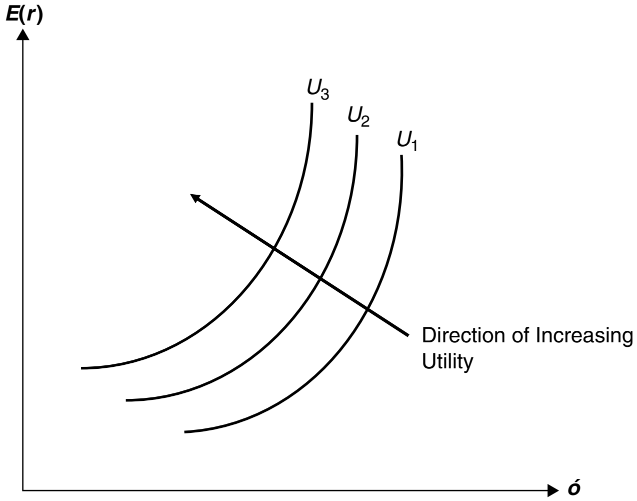


**EXHIBIT 2.10** Risk and Return Indifference Curves

Let's repeat this in a slightly different way. Starting on the left where σ and *E*(*r*) are both low, a relatively small increase of *E*(*r*) will just compensate the representative investor for a small increase in risk. As you move further to the right along the same indifference curve, note that successively larger increases in *E*(*r*) will be required for the representative investor's utility (of wealth) to remain constant as risk is successively increased by small increments of equal size. Alternatively stated, as we move to the right along a rising indifference curve, successively larger increases in *E*(*r*) are required because the investor's marginal risk premium is going up.

In our previous analysis, the slope of the *X*, *Y* indifference curve was (the negative of) the marginal utility of *X* divided by the marginal utility of *Y*, and the slope of the current consumption, future consumption indifference curve was (the negative of) the marginal utility of current consumption divided by the marginal utility of future consumption. In the current analysis, the slope of the risk, expected return indifference curve reflects the marginal *dis*utility of risk (risk is a bad) divided by the (positive) marginal utility of a risk premium. The slope of an indifference curve is the negative of the ratio of marginal utilities, and the marginal utility of risk is itself negative; thus the slope of a risk and return indifference curve is positive.

As we move along an upward sloping risk, expected return indifference curve, the marginal disutility of risk increases, the marginal utility of expected returns declines, and the ratio of the two marginal utilities increases. This translates into the decision maker's marginal risk premium increasing along an indifference curve, moving from left to right. That is why the indifference curves in Exhibit 2.10 are convex from below.

### Optimal Solution to the Risk-Return Trade-Off Problem

The risk-return trade-off is resolved with reference to the investor's underlying objective. For a consumer choice model under certainty, the objective is the maximization of utility. Under risk, as we have noted, the objective becomes the maximization of *expected utility*. Otherwise the methodology stays the same.

As in the previous two sections, we bring the budget constraint (here the capital market line) and the indifference curve mapping together to see how our representative investor maximizes expected utility by selecting an optimal risk and expected return combination. The solution, which is shown in Exhibit 2.11, is given by the point labeled *A* where the indifference curve labeled $E(U)^*$ is tangent to the capital market line. Moving toward the northwest, $E(U)^*$ is the highest indifference curve that the investor can reach. On the vertical axis in Exhibit 2.11, the point labeled $E(r)^*$ is the optimal amount of expected return to select; on the horizontal axis, the point labeled $\sigma^*$ is the optimal amount of risk to accept.

Identifying the optimal values for risk and expected return [$\sigma^*$ and $E(r)^*$] establishes the optimal portfolio weights for the risky market portfolio and the riskless asset. To see this, go back to equation (2.13): $w_M = \sigma/\sigma_M$. If, for instance, $\sigma_M = .08$ and $\sigma^* = .04$, than $w_M{}^* = 0.5$ and the investor maximizes expected utility by putting half of his or her assets in the risky market portfolio and half in the riskless asset.

### Interpretation

When the investor is holding an optimal combination of the risky market portfolio and the riskless asset, it can be shown that[7]

$$MRP = \frac{E(r_M) - r_f}{\sigma_M} \tag{2.15}$$

---

[7] For a proof and further discussion, see Robert Schwartz, *Equity Markets: Structure, Trading, and Performance* (Harper & Row, 1988) and Reto Francioni, Sonali Hazarika, Martin Reck, and Robert A. Schwartz, "Security Market Microstructure: The Analysis of a Non-Frictionless Market," *Handbook of Quantitative Finance*, eds. Cheng-Few Lee and Alice C. Lee (Springer, 2010).

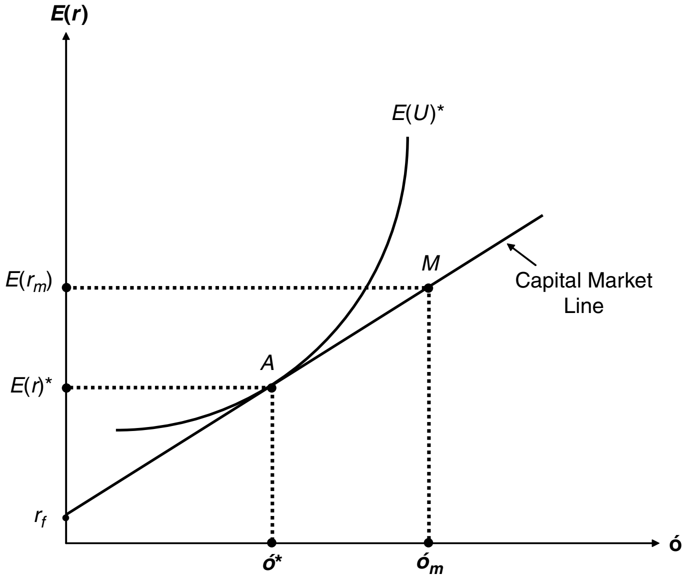


**EXHIBIT 2.11** Determination of the Decision Maker's Optimal Risk and Expected Return Combination

where *MRP* is the investor's marginal risk premium per unit of risk. This is an insightful statement of the investor's marginality condition: The optimal holding of the risky asset is the number of shares that equates the investor's *MRP* (per unit of risk in the investor's portfolio) with the price of risk that the market will pay (per unit of risk in the market portfolio).

We have seen the right-hand side of equation 2.15 before in this section of the chapter. It is the slope of the capital market line. In equilibrium, the capital market line is tangent to the investor's indifference curve. It is now clear how our representative investor's marginal risk premium relates to the price of risk. When the investor is holding an optimally structured portfolio, the marginal risk premium for that portfolio and the market price of risk are brought into equality.

With equality achieved, the investor has no further desire to trade in the marketplace for these two financial assets. Moreover, as all participants achieve this equality, there is no longer any incentive for anyone to trade with anyone else. The frictionless, competitive marketplace will have served all of the investors very well indeed.

Once again we see the power of a micro market to deliver a highly efficient outcome in a frictionless environment where no individual participant has any market power over price.

## COMPARATIVE STATICS

In the preceding three sections of this chapter, we modeled the consumer's optimal decision in a trade-off situation, given his or her income (or wealth), tastes, and market prices. This is *static analysis*. With all the independent variables constant, we obtained a single equilibrium point.

We now proceed to *comparative statics*. This step involves changing one of the independent variables (e.g., a relevant price or income) while holding all else constant and seeing how the dust settles. Comparative statics involves contrasting the two equilibrium points. *Dynamic analysis* (which we do not get into here) would take us one step further: modeling the *adjustment path* that takes the consumer from one equilibrium to another.

We now return to each of the three trade-offs that we considered in the preceding three sections, and focus on just one thing: a change in the relevant price variable. This step can easily be extended so that we might obtain a consumer's demand curve for a good, service, or asset. We focus on the demand curve in Chapter 3.

### The X, Y Trade-Off

Exhibit 2.12 displays the comparative statics for the trade-off between our two undefined goods, $X$ and $Y$. In the diagram, $P_Y$ (the price of Y), $M$ (income), and tastes (the indifference curves) are held constant, while $P_X$ (the price of $X$) is assigned two different values, $P_{X1}$ and $P_{X2}$, with $P_{X1} > P_{X2}$. The top diagram in the exhibit presents the budget constraint and indifference curve mapping. The bottom diagram makes the comparative statics explicit. In the bottom diagram, the price of $X$ is on the vertical axis and the optimal amount of $X$ to purchase in the market is on the horizontal axis (as it is in the top diagram).

As shown in Exhibit 2.12(a), as the price of $X$ decreases from $P_{X1}$ to $P_{X2}$, the budget constraint rotates out to the right. As it does, the $Y$ intercept remains fixed at $M/P_Y$ (in keeping with our *ceteris paribus* tradition, both $M$ and $P_Y$ are constant) while the $X$ intercept shifts to the right from $M/P_{X1}$ to $M/P_{X2} > M/P_{X1}$.

In Exhibit 2.12(a), the first budget constraint is labeled (1) and the second is labeled (2). The slope of constraint (2) is clearly a smaller negative number (that is, smaller in absolute value) than the slope of constraint (1). Consequently, the second constraint is tangent to an indifference curve further to the right along the curve, where the indifference curve's slope is also a smaller negative number. Importantly, further to the right along an indifference curve, a larger quantity of $X$ is consumed.

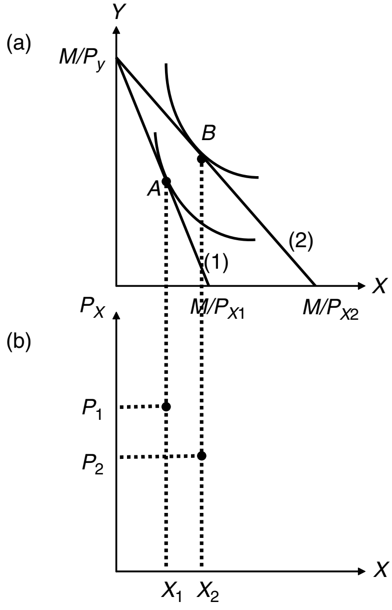


**EXHIBIT 2.12** A Decision Maker's Comparative Statics Regarding an Optimal $X$, $Y$ Combination [2.12(a)] and Optimal Consumption of $X$ [2.12(b)]

In Exhibit 2.12(a), points $A$ and $B$ identify the two optimality solutions for the consumer. We have dropped vertical lines from points $A$ and $B$ in the top diagram down to the horizontal axis of the lower diagram to show the two different quantities of $X$ obtained at the two different prices of $X$. Exhibit 2.12(b) shows that a price of $X$ of $P_1$ is matched with $X_1$ units of $X$ being demanded. We also see that a price of $P_2$ is matched with $X_2 > X_1$ units being demanded. Clearly, more units are demanded at the lower price: $X_2$ is greater than $X_1$ with $M$, $P_Y$, and tastes constant. This is the comparative statics that in the next chapter will take us to the downward sloping demand curve for the undefined good $X$.

Notice that the rotation of the budget constraint around a fixed point on the $Y$ axis (a rotation that has swung the budget constraint out and to the right) has enlarged the opportunity set for the consumer and, in so doing, has enabled the consumer to reach a higher indifference curve. We have held the individual's nominal income ($M$) constant, and have effectively increased his or her *real income* (that is, command over real resources). Consequently, the difference in the amount of $X$ sought for purchase in the market ($X_2 - X_1$) can be decomposed into two parts:

1. **A substitution effect**, which is attributable to a move along a convex indifference curve. That is, with $P_{X2} < P_{X1}$, the marginal rate of substitution of $X$ for $Y$ has to be a smaller negative number at the new tangency point, and this is achieved by substituting $X$ for $Y$, that is, by moving down and to the right along a given, convex indifference curve to a point where the $MRS_{X.Y}$ is indeed a smaller negative number.
2. **An income effect** that is attributable to a move to a higher indifference curve.

The decomposition is illustrated in Exhibit 2.13. The exhibit displays three linear budget constraints (labeled 1, 2, and 3) and two convex indifference curves (labeled $U_1$ and $U_2$). The three budget constraints have the following $Y$ and $X$ intercepts, respectively:

1. **Budget constraint 1:** $M/P_Y$, $M/P_{X1}$
2. **Budget constraint 2:** $M'/P_Y$, $M'/P_{X2}$
3. **Budget constraint 3:** $M/P_Y$, $M/P_{X2}$

**The Substitution Effect:** Budget constraints 1 and 2 are both tangent to the lower indifference curve, $U_1$. Budget constraint 2 reflects the lower price

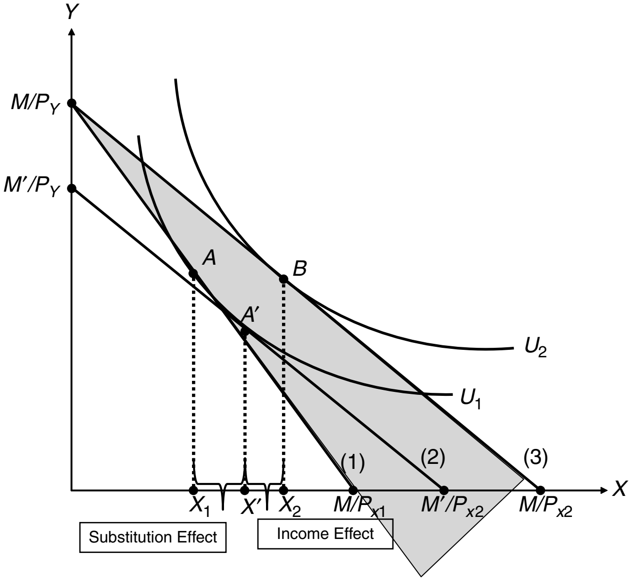


**EXHIBIT 2.13** Decomposition of a Quantity Response to a Price Change: An Income Effect and a Substitution Effect

of $X$ (which is $P_2$) and thus, with $P_Y$ the same, the slope of budget constraint 2 is a smaller negative number (smaller in absolute value) than the slope of budget constraint 1. Consequently, budget constraint 1 is tangent to $U_1$ at point $A$, and budget constraint 2 is tangent to $U_1$ at point $A'$, with $A'$ being below and to the right of $A$. Thus the optimal quantity of $X$ at point $A$ is $X_1$, while at $A'$ the optimal quantity is $X'$. The difference, $X' - X_1 > 0$, is the pure substitution effect.

**The Income Effect:** Budget constraints 2 and 3 both reflect the lower price of $X$ (which is $P_2$), and therefore, with $P_Y$ constant, they are parallel. However, budget constraint 2 reflects a lower nominal income of $M'$ than does budget constraint 3, which is based on an income of $M$. The difference, $M - M' > 0$ can be understood as a decrement in nominal income that would keep the household's real income (utility) constant despite the price of $X$ being lower while the price of $Y$ is constant. Budget constraint 2 is tangent to $U_1$ at point $A'$, and budget constraint 3 is tangent to $U_2$ at point $B$, with $B$ being above and to the right of $A'$. The optimal quantity of $X$ at $A'$ is $X'$, while the optimal quantity of $X$ at B is $X_2$. The difference, $X_2 - X' > 0$ is a pure income effect. Household consumption rises for most goods as income rises and, accordingly, the income effect shown in Exhibit 2.13 is positive.

This completes the comparative statics for the $X$, $Y$ trade-off.

## The Current Consumption, Future Consumption Trade-Off

Exhibit 2.14 displays the comparative statics for the present consumption, future consumption trade-off. In the diagram, $Y_0$ and $Y_1$ (current income and future income, respectively) and tastes (the indifference curves) are held constant, while the price of current consumption relative to the price of future consumption has two different values, $(1 + r_1)$ and $(1 + r_2)$, with $(1 + r_1) > (1 + r_2)$. The top diagram in the exhibit shows the two budget constraints and the indifference curve mapping. The bottom diagram makes the comparative statics explicit: The price of current consumption relative to future consumption is on the vertical axis, and the optimal amount of current consumption, $C_0$, is on the horizontal axis.

As shown in Exhibit 2.14(a), as the price relative decreases from $(1 + r_1)$ to $(1 + r_2)$, the budget constraint rotates counterclockwise around the fixed point, $Y_0$, $Y_1$. In Exhibit 2.14, the budget constraint that reflects the higher price is labeled $(1 + r_1)$ and the constraint that reflects the lower price is labeled $(1 + r_2)$. The slope of the budget constraint labeled $(1 + r_2)$ is clearly a smaller negative number (smaller in absolute value) than the slope of the budget constraint labeled $(1 + r_1)$. Consequently, the tangency

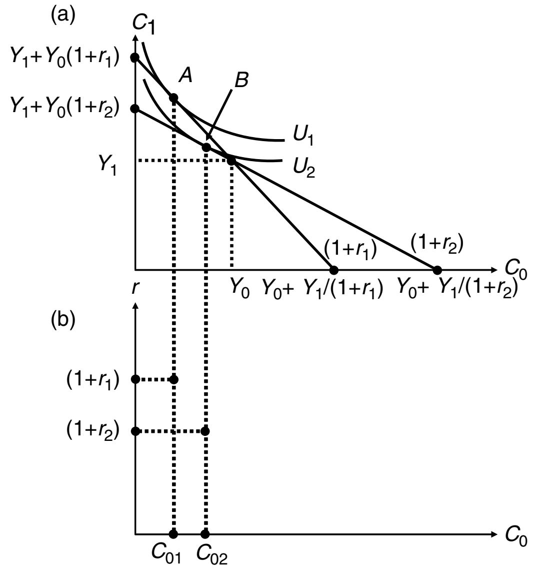


**EXHIBIT 2.14** A Decision Maker's Comparative Statics Regarding an Optimal Current Consumption, Future Consumption Combination [2.14(a)] and Optimal Amount of Current Consumption [2.14(b)]

point for the second constraint is further to the right along the indifference curve it is tangent to ($U_2$) than is the case for the first constraint, which is tangent to the $U_1$ indifference curve.

Points *A* and *B* in Exhibit 2.14(a) identify the tangency points for the higher and the lower price relatives, respectively. These points show that further to the right along an indifference curve (where the curve's slope is flatter), current consumption is greater; that is, current consumption is greater at the lower price than it is at the higher price. This result should come as no surprise, but it is important to have derived it analytically.

As in Exhibit 2.12, we have dropped vertical lines in Exhibit 2.14(a) from points *A* and *B* in the top diagram, down to the horizontal axis of the lower diagram where we match the two different values of current consumption with the two different price relatives. Exhibit 2.14(b) shows that if the price relative is $(1 + r_1)$, then current consumption is $C_{01}$. However, if the price relative is $(1 + r_2) < (1 + r_1)$, then current consumption is $C_{02} > C_{01}$. This analysis formalizes the fact that, all else constant, current consumption is greater when interest rates are lower. This certainly makes intuitive sense: With lower interest rates, borrowing costs are less (it is easier to increase current consumption by borrowing) and the market does not pay as much to lenders (it is more difficult to increase future consumption by lending).

As with the $X$, $Y$ trade-off, an income effect as well as a substitution effect attends the change in the price of current consumption relative to future consumption. The substitution effect is similar to the one that attends a standard $X$, $Y$ trade-off: for instance, if a decrease of the interest rate lowers the price of present versus future consumption, the consumer will substitute present consumption for future consumption either by lending less or by borrowing more. The income effect that attends the current consumption, future consumption trade-off, however, is a bit trickier. For the $X$, $Y$ trade-off, the income effect generated by a decrease in $P_X$ (all else constant) is unambiguously positive: The budget constraint, by rotating in a counterclockwise direction around a fixed point on the $Y$ axis, has enabled the household to reach a higher indifference curve ($U_2 > U_1$).

For the current consumption, future consumption trade-off, the budget constraint also rotates in a counterclockwise direction as the relative price of current consumption decreases; however, it does so around a point that is not on the vertical axis, but is internal to the diagram, the point that is set by the configuration of the income stream, $Y_0$, $Y_1$. As is shown in Exhibit 2.14, the $(1 + r_2)$ budget constraint is tangent to a lower indifference curve than is the $(1 + r_1)$ budget constraint (i.e., we have $U_2 < U_1$). The reason is that both tangency points ($A$ and $B$) involve saving that enables more consumption in the next period, and a lower interest rate means that next year's consumption will not be higher by as much. You might also notice that, to the left of the $Y_0$, $Y_1$ point, the $(1 + r_2)$ budget constraint is below the $(1 + r_1)$ budget constraint, which means that the obtainable set of consumption baskets is more restricted in this region when interest rates are lower.

On the other hand, the set of obtainable consumption baskets is greater to the right of the $Y_0$, $Y_1$ point, for in this region the $(1 + r_2)$ budget constraint is above the $(1 + r_1)$ budget constraint. If for both interest rates the household achieves optimality in this region, it is borrowing in order to spend more now, and it will consume less in the future. Accordingly, a lower interest rate enables any given increase of current consumption to be attended by a smaller decrement of future consumption, and thus a higher level of utility can be attained. You might assure yourself of this result by drawing in steeper indifference curves that reach tangency to the right of the $Y_0$, $Y_1$ point.

It is also possible to have two budget constraints that are both tangent to a single indifference curve. In this case, the household will lend at the higher interest rate and borrow at the lower interest rate, and no income effect at all will attend this particular interest rate change. We suggest that you take out a clean piece of paper and try drawing the two budget constraints and indifference curve mapping that give this result.

This completes the comparative statics for the current consumption, future consumption trade-off.

## The Risk-Return Trade-Off

Exhibit 2.15 displays the comparative statics for the risk, return trade-off. In the diagram, $\sigma_M$ (the riskiness of the market portfolio) and tastes (the indifference curves) are held constant, while the expected return on the market has two different values, $E(r_{M1})$ and $E(r_{M2})$, with $E(r_{M2}) > E(r_{M1})$. The top diagram in the exhibit shows the two capital market lines and the indifference curve mapping. The bottom diagram makes the comparative statics explicit: In the bottom diagram, $E(r_M)$ is on the vertical axis and the *optimal amount* of $\sigma$ to accept is on the horizontal axis.

As shown in the top diagram, as $E(r_M)$ increases from $E(r_{M1})$ to $E(r_{M2})$, the capital market line rotates counterclockwise around a fixed point, the vertical intercept, $r_f$. In Exhibit 2.15(a), the capital market line that reflects

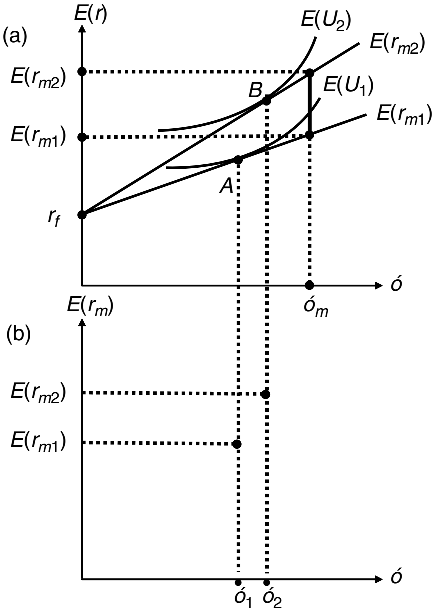


**EXHIBIT 2.15** A Decision Maker's Comparative Statics Regarding an Optimal Risk and Expect Return Combination [2.15(a)] and Optimal Amount of Risk for Two Diff-erent Levels of Expected Market Return, $E(r_M)$ [2.15(b)]

the lower expected return is labeled $E(r_{M1})$, and the capital market line that reflects the higher expected return is labeled $E(r_{M2})$. The slope of the $E(r_{M2})$ line is clearly a larger positive number than the slope of the $E(r_{M1})$ line and, consequently, it is tangent to the higher indifference curve $[E(U_2)]$ and the point of tangency (labeled $B$) is further to the right along the curve than is the case for the lower indifference curve $[E(U_1)]$ which is tangent to the $E(r_{M1})$ budget constraint at point $A$. Points $A$ and $B$ in Exhibit 2.15(a) identify the tangency points for the lower and the higher capital market lines, respectively. These points show that, further to the right along an indifference curve where the curve's slope is steeper, a larger value of $\sigma$ is optimal when the expected return is $E(r_{M2}) > E(r_{M1})$.

As in Exhibits 2.12 and 2.14, we have dropped vertical lines in Exhibit 2.15(a) from points $A$ and $B$ in the top diagram, down to the horizontal axis of the diagram in Exhibit 2.15(b) where we contrast the two values of $\sigma$ (which are $\sigma_1$ and $\sigma_2$) that are shown on the horizontal axis, with the two values of $E(r_M)$ [which are $E(r_{M1})$ and $E(r_{M2})$] that are shown on the vertical axis. Now let's focus on Exhibit 2.15(b), where $E(r_M)$ is shown explicitly on the vertical axis. In Exhibit 2.15(b), we see that, if the expected return is $E(r_{M1})$, then the risk taken on by the investor is $\sigma_1$. However, if the expected return is $E(r_{M2}) > E(r_{M1})$, then the risk taken by the investor is $\sigma_2 > \sigma_1$. This analysis formalizes the fact that, all else constant, risk taking is greater when expected returns are higher. This certainly makes intuitive sense: a higher risk premium (a premium that the market pays the investor) encourages the risk averse investor to accept more risk).

As with the $X$, $Y$ trade-off, an income effect as well as a substitution effect attends the change in the slope of the capital market line. As $E(r_M)$ increases, the capital market line, by rotating in a counterclockwise direction around a fixed point on the vertical axis, enables the investor to reach a higher indifference curve $[E(U_2) > E(U_1)]$.

Unlike our analyses of the $X$, $Y$ trade-off and of the current consumption, future consumption trade-off, we have not yet made two key variables explicit with regard to the risk, return trade-off: the price of a share of the risky market portfolio, and the number of shares that the investor will wish to hold. Let's take that step now. It will move us closer to obtaining a downward sloping demand curve to hold shares of the risky market portfolio.

To proceed, think in terms of a simple, single holding period model. That is, the investor acquires shares of the market portfolio now (call this point in time 0) and holds them for one period, at which time (call it point in time $T$) the holdings are liquidated and the proceeds are redeployed (perhaps, in keeping with our previous one period analysis of the current consumption, future consumption trade-off, they are spent on consumption goods). Let $P_0$ denote the current share price of the market portfolio and let

$E(P_T)$ denote the market portfolio's expected future share price. Consistent with our earlier discussion of returns measurement, we define the expected return for the holding period of length $T$ as

$$1 + E(r_M) = \frac{E(P_T)}{P_0} \tag{2.16}$$

The right-hand side of Equation 2.16 is a *price relative*. Multiplying both sides by $P_0$ puts the equation in a form that you might be more familiar with: $E(P_T) = P_0[1 + E(r)]$. Log returns are typically defined with regard to the price relative, a variable that can never be negative (which is a necessary requirement because it is not possible to take the logarithm of a negative number). We will not, however, encumber the discussion at this point by taking logs.

Equation 2.16 makes the current share price of the market portfolio explicit. It shows that, $E(P_T)$ constant, $E(1 + \mathrm{r}_M)$ is a decreasing function of $P_0$, the current price of a share of the risky market portfolio. For our purposes here, we in fact hold $E(P_T)$ constant (i.e, it is a *ceteris paribus* variable). We do so because we are focusing on how the current share price is set *given* our expectations of what the future share price will be at time $T$. It is clear that, given $E(P_T)$, our expected return will be higher if shares can be purchased today at a lower price.

Take another look at Exhibit 2.15(b). In that lower diagram, it shows that a larger value of $E(r_M)$ is associated with a larger value of $\sigma$, as the increase in the expected return shifts the point of tangency further out on the higher capital market line [as shown in Exhibit 2.15(a)]. Further, a move to larger values of $E(r_M)$ and $\sigma$ on the capital market line is accomplished by increasing the weight of the risky market portfolio in the investor's combined portfolio. And how is this weight increased? It is raised by increasing the number of shares held of the risky market portfolio ($N_M$) at a rate greater than the rate at which the per share price ($P_M$) has decreased (recall that the weight is $w = P_M N_M / W$).

Putting this all together, we achieve our final result: *ceteris paribus* (i.e., holding constant the risk-free rate of interest [$r_f$], the expected future share price [$E(P_T)$], and tastes for risk), the investor will hold more shares of the risky market portfolio at a lower current price per share. This relationship is shown in Exhibit 2.16 where the current share price is on the vertical axis and the number of shares held is on the horizontal. At point $A$, a current share price of $P_{01}$ is matched with a share holding of $N_1$, and at point $B$ a lower share price of $P_{02}$ is matched with a larger share holding of $N_2 > N_1$.

This completes the comparative statics for the risk, return trade-off. Having considered three different trade-off situations, we are now in a

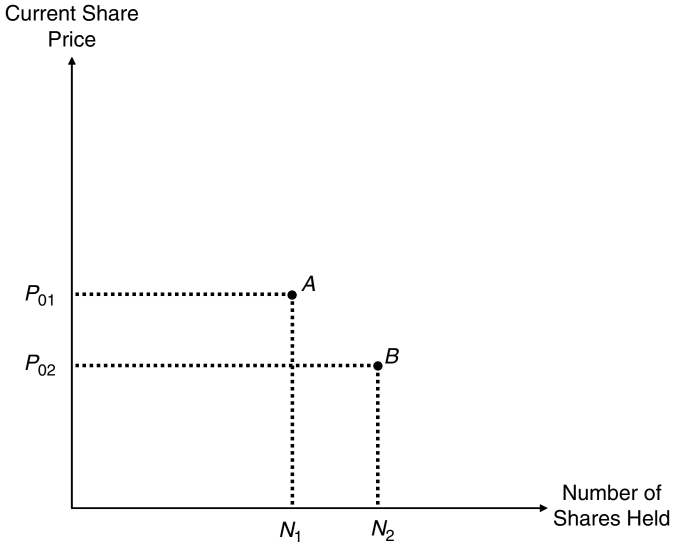


**EXHIBIT 2.16** Comparative Statics Regarding the Number of Shares Held for Two Alternative Values of Share Price

good position to move forward and deal with demand curves. We do so in the next chapter.

## CHAPTER SUMMARY

The chapter has presented the consumer choice model with regard to three different trade-offs that each involve two items: (1) a trade-off between generic good or service *X* and generic good or service *Y*, (2) a current consumption versus future consumption trade-off, and (3) the risk-return trade-off that plays a central role in modern portfolio theory. As you will appreciate in Chapter 3, with the consumer choice model in hand we can derive a decision maker's demand curve for a good or service, or for a financial asset.

Key thoughts developed in this chapter that you should have a solid grasp of include:

- To resolve optimally the two-item trade-offs, two relationships were specified for each: (1) a budget constraint that demarks the "obtainable set" (the alternative achievable combinations of the two-items), and

(2) a family of indifference curves that can be derived from the decision maker's utility function which describes his or her tastes.

- When the decision maker has resolved the trade-offs optimally (that is, has selected a combination that maximizes his or her utility or expected utility), his or her tastes are harmonized (on the margin) with the realities of the marketplace. Specifically, when the decision maker has achieved an optimal combination of the two factors, the ratio of relative prices (as given by the slope of the budget constraint) equals the ratio of marginal utilities (as given by the slope of the highest indifference curve that he or she can reach).
- For the *X*, *Y* trade-off and the present consumption, future consumption trade-off, both items are "a good" and the decision maker's indifference mappings for these goods are negatively inclined. For the risk-return trade-off, expected return is "a good"—the decision market's expected utility is higher when the expected return is higher (risk constant), but risk is "a bad"—the decision market's expected utility is lower when the risk is higher (expected return constant). This is because the decision maker is assumed to have a risk-averse utility function for income (or wealth). With return a good and risk a bad, risk-return indifference curves are positively inclined.
- The decision maker's risk aversion is not attributable to any attitude regarding the act of risk taking per se (i.e., the fun or displeasure that might be derived from playing a risky game), but to the decision maker's utility of wealth increasing at a diminishing rate. Consider a betting situation that is fair in monetary terms. A bet is fair if the expected net monetary payout from taking the bet is zero. With the utility of income (or of wealth) decreasing on the margin, the utility gained from winning a bet that is fair in monetary terms is less than the utility that is lost when the fair bet is lost.
- That is, if a bet offers you an equal chance of winning or of losing a dollar, the utility you would gain from the dollar you might win is less (remember, your marginal utility of dollars is decreasing) than the utility you would lose by losing a dollar. Thus your expected utility gain from taking the bet is negative. Consequently, in utility terms the bet is not fair, which means that a positive expected monetary return must be offered if a risk-averse participant is to accept taking the risk.
- To analyze a risk-return trade-off for an investor, a stock's (or portfolio's) expected return was given precise definition. Strictly speaking, it is the expected logarithm of price relatives, with prices having been adjusted for cash and stock dividends, and for stock splits. Risk was defined as the variance of logarithmic returns, assuming that returns

are lognormally distributed. For simplicity, however, we suppressed the label "logarithm" and dealt with arithmetic returns as if they are normally distributed.

- We have noted that a risk-averse investor will accept a risky situation only if compensated for taking the risk. The compensation is in the form of a higher expected return, and the additional compensation that is required to induce the investor to take the risk is a "risk premium."
- With the four key terms identified (risk aversion, expected return, risk, and risk premium), the risk-return trade-off was analyzed assuming two financial assets: (1) a risky market portfolio (e.g., a well-diversified mutual fund) and (2) a risk-free asset (e.g., Treasury bills). Working in this setting we obtained (1) a budget constraint that we called the "Capital Market Line" and (2) a family of positively inclined risk-return indifference curves. These curves, and the optimality solution that they delivered, are consistent with the Capital Asset Pricing Model (CAPM), a cornerstone of modern portfolio theory.
- The preceding equilibrium models were extended to show how an optimal selection of a generic good ($X$), or a more defined good (current consumption), or the acceptance of a characteristic (risk) changes when the price of the good or characteristic changes, all other independent variables constant. These comparative static results will be generalized in Chapter 3 to give us the decision maker's demand curve for each good or characteristic.

## LET'S CONSIDER AN APPLICATION: PHILANTHROPY

Are the following three statements mutually consistent?

1. In microeconomic theory, utility maximization is the goal of a representative consumer (household).
2. Utility is a function of the goods and services that the decision maker consumes.
3. Some people give money, time, goods, and services to others and to charitable organizations.

Let's think about this one. If your goal is to maximize your utility, if your utility is a function of your personal consumption, and because the amount you can consume is greater when your income is greater, why give your income away? Is the act of giving irrational economic behavior? Or are we being too restrictive by implying that your utility is a function only of your own personal consumption? Perhaps. Let's bring charity into the picture.*

The more formal term for charity is *philanthropy* (which translated from the Greek means "love of man"). But not all giving is philanthropic. Giving might also be motivated by a desire for self-promotion, a kind of personal advertising expenditure, a statement that says, "I am a good person, I am an important person." Nevertheless, some people are concerned about the welfare of others and they give because they care. Can our utility-theoretic formulation handle this?

Incorporating philanthropic behavior into the utility-theoretic framework could not be simpler. We do so by introducing the concept of an *interrelated utility function*. That is, your utility is a function not only of your own consumption, but also of the consumption of others, because you care about the welfare of others. This can be formalized.

In the chapter we analyzed consumer choice by keeping it simple, by working in a two-product (or product characteristic) world ($X$ and $Y$, current consumption and future consumption, risk and return). Let's continue to do this. Simply divide all consumption into two categories: the amount that the philanthropist consumes

---

*Material for this application draws from Robert A. Schwartz, "Personal Philanthropic Contributions," *Journal of Political Economy* (November/December 1970), 1264–1291.

and the amount that a recipient of the charitable act consumes. Call the philanthropist A and the recipient B. Then A's utility function can be written as

$$U^{A} = U^{A}\left(C^{A}, C^{B}\right)$$

where $C^{A}$ and $C^{B}$ are A's and B's consumption, respectively.

With this by way of background, we invite you to do the following.

1. Construct an indifference curve mapping for household A that has A's income and consumption on the horizontal axis, and B's income and consumption on the vertical axis.
2. Add a budget constraint to your diagram. To do so, let $Y^{A}$ and $Y^{B}$ be A's and B's income, respectively, and use $Y^{A}$ and $Y^{B}$ to locate the height of the constraint. The slope of the budget constraint is the negative of the price to A of a dollar's worth of personal consumption relative to the price to A of increasing B's consumption by one dollar. If A giving away a dollar results in B receiving a dollar, the slope of the budget constraint is –1.
3. Let $D^{AB}$ be A's *donation* to B. Use your diagram to identify A's equilibrium amount of personal consumption and the optimal gift, $D^{AB}$, that A should give to B.
4. Note the similarity between the diagram that you have just produced and the present consumption, future consumption diagram in Exhibit 2.5.
5. Use your diagram to verify that $D^{AB}$ is a positive function of $Y^{A}/Y^{B}$.
6. Present an indifference curve mapping which shows that if $Y^{A}/Y^{B}$ is small enough A would be better off receiving a gift from B, even though for $Y^{A} \geq Y^{B}$ A's attitude toward B, on the margin, is philanthropic.
7. Let $T_{M}$ be A's marginal tax rate and assume that A's gift to B is tax-deductible (as some giving is). How is A's budget constraint affected? What affect does tax deductibility have on giving rates? What effect does an increase in tax rates have on giving rates?

## TraderEx SIMULATION: WHAT ARE YOUR ATTITUDES TOWARD RISK?

The chapter has paid substantial attention to the concept of risk aversion as it relates to a decision maker's utility function for income (or wealth), and to risk and return indifference curves. Investing in equity shares certainly is risky and, the greater the risk, the greater the expected return that a risk-averse investor will require before acquiring an equity position. Does the link between diminishing utility of income (or wealth) and risk aversion seem like an abstract idea to you? Perhaps you can get a better sense of your own attitude toward risk by using the TraderEx simulation.

The software that comes with *Mastering the Art of Equity Trading Through Simulation* has two assets: a risky equity asset and cash (the risk-free asset). In the current exercise, we will start you off with \$10,000 of cash in your simulated account and zero shares of a risky asset that you can buy with the \$10,000. The simulation that you will run has two trading days that are separated by one year. On "day one," which is at the beginning of the year, you are a buyer. By buying, you establish a position that you will hold for one year. On "day two," which is at the end of the year, you are a seller. By selling, you liquidate your position and turn your shares back into cash. Hopefully you will have more cash at the end of day two than you had at the beginning of day one.

Your ending cash position divided by your opening cash position, minus one, is the rate of return that you will have realized for the simulation run. Once the simulation parameters have been set, the stock's one-year return from the end of day one to the start of day two is out of your control. The computer gives you this return by randomly drawing from a lognormal distribution that has a mean (the expected return) and variance (the risk). You will see this one-year return virtually instantaneously.

You will have a five-minute period (clock time) on day one to establish your equity position, and a five-minute period on day two to trade out of your stock position and see how well you have done. Trade carefully on day one when you put on your position and on day two when you liquidate your position. If you buy at relatively high prices on day one and sell at relatively low prices on day two, your one-year return will be lower. In this exercise, however, we focus on the investment return and deemphasize the effect that trading costs can have on the longer period return. We achieve this focus in the exercise by giving you very short trading periods (just five minutes of clock time) and a highly liquid market (a speedy execution of a \$10,000 order should have a minimal impact on your one-year return). In the end of chapter material for Chapter 4, we change the emphasis from the investment component to the trading component by giving you a similar simulation structure but longer trading periods and considerably thinner books (books with fewer orders on them).

For the current exercise, at the start of each simulation you can set two parameters: the expected return [E(R)] and the standard deviation [SD(R)]. Do so using four of the following eight combinations (your choice):

| E(R) | SD(R) |
|---|---|
| .05 | .20 |
| .10 | .15 |
| .15 | .10 |
| .20 | .05 |
| .05 | .05 |
| .10 | .10 |
| .15 | .15 |
| .20 | .20 |

For each of the four combinations that you have chosen, run the simulation enough times so that you get some high returns (above the mean) and some low returns (below the mean). Then do the following:

1. Make note of the high return, the low return, and the average return realized for each combination, and rank the combinations according to your preference concerning the risk, return properties. For example, would you prefer an asset with a high average return but that occasionally delivers a sizable loss (i.e., a high expected return, high standard deviation asset), or would you prefer a relatively stable asset with a relatively low average return (i.e., a low expected return, and low standard deviation asset).
2. For each of the four combinations, explain whether the expected return was high enough (in your opinion) to compensate you for the risk involved.
3. If you are particularly satisfied or dissatisfied with a specific risk-return combination, explain why.
4. Try a couple of other E(R), SD(R) combinations in an attempt to locate two combinations that you would be indifferent between (that is, which would put you on the same indifference curve).
5. Do you now understand better how your attitude toward risk is related to your utility function for income (or wealth)?

# CHAPTER 3

# Demand Meets Supply

In this chapter we extend the comparative statics presented in Chapter 2 to obtain a demand curve, which is the functional relationship between the price of an economic good and the quantity of that good which is demanded. We first deal with the individual household's demand for the generic product $X$, and then show how we use the measure of responsiveness called *elasticity* to represent the sensitivity of one variable, the quantity of $X$ demanded, to, respectively, the price of $X$ ($X$'s own price elasticity of demand, $P_Y$ and income constant), the price of the other generic product $Y$ ($X$'s cross-price elasticity of demand, $P_X$ and income constant) and the decision maker's income ($X$'s income elasticity of demand, $P_X$ and $P_Y$ constant).

The next step involves aggregating the decisions of the individual participants so that we might understand the implications of their collective behavior. Aggregation (a simple summation process) involves making the transition from individual demand curves to the market's aggregate demand curve. With the aggregate market demand curve in hand, we introduce a supply curve and show how the intersection of demand and supply in a free and competitive market establishes equilibrium values for price and quantity, our two key variables.

We use the demand-supply formulation to show the effect of regulatory intervention on market equilibrium (specifically, we consider the imposition of a price control, a sales tax, and a subsidy). The demand formulation and equilibrium analysis is presented for the generic good, $X$, and for a risky financial asset. As we proceed with the discussion, we continue in this chapter to invoke the simplicity of a frictionless market.

But before we get ahead of ourselves, let's back up and review the decisions of a representative consumer.

## SHOULD HAVE, WOULD HAVE, AND DID: THE DECISIONS OF A REPRESENTATIVE CONSUMER

A market is driven by the decisions of all of the individuals who interact within it, and the first step to understanding a market's dynamics is to comprehend the decisions of its participants. We referred any number of times in Chapter 2 to the decisions of a *representative consumer*. By "representative," we mean the central tendency of a larger group of participants. In other words, we are referring to the average.

Of course not everybody behaves in accordance with our representative consumer. Following a jump in the price of gas, for instance, some households might be observed using their cars more, not less. Our number-one interest is not in a specific individual per se but, as we just said, in the average, in how drivers in aggregate respond to the market forces that are exerted by a higher price of gas at the pump. Does the *representative household* cut back on its driving? That is what the consumer choice model predicts should happen.

But "should" is a tricky little word. It has two meanings. If our model says that a consumer should consume less gas if the price goes up, "should" is being used as a prediction. The thought can be expressed equally as well by saying, "we *predict* that less gas will be consumed when the price of gas is higher, all else constant." Closely aligned with prediction is the desire to understand. What is it that makes a clock tick, what causes water to boil, what leads a free market to deliver *socially desirable results*? With respect to the latter, the consumer choice model sheds light on the answer.

"Should," however, can also be used as a prescription, and it often is. When giving advice, we might use the word in sentences such as the two that follow: "From all that you are saying, I think that what you *should* do is . . . I am not predicting that you will follow my advice, but this is what I really think you *should* do. . . ."

A microeconomic model that is seeking to predict behavior and/or to understand behavior is a *positive model*. In positive analysis, "should" is used in the predictive sense. In contrast, a microeconomic formulation designed to give advice to a client is a *normative model*. In normative analysis, "should" is used in the sense of an instruction or a command. The difference in use does not call for a totally different formulation, however. To the contrary, positive models can provide very good foundations for normative models, and a prescription may, at heart, be based on a prediction: "I advise that you do this. In fact you *should* (a normative "should") do this, because, if you do so, the following *should* (a positive "should") happen." Nevertheless, be careful. The literal application of a microeconomic model to an actual decision maker (be it a government agent, a corporate client, or even a family member) may call for further elaboration and more precision than does the positive model that it is based on. Moreover, a formulation that may not work well in a specific normative context can still have considerable value in a positive application.

Our primary focus in this book is on the positive use of economic analysis. We are seeking to predict and to understand how markets behave. There may, however, be some normative applications of the analysis further down the road. One use could involve the formulation of good public policy. If it is our understanding and our prediction that the free market will produce socially desirable results, then government should (a normative "should") avoid undue regulatory intervention in the operations of the free market. On the other hand, if for well-understood reasons the free market fails to produce socially desirable results, and if a source of market failure can be identified (we noted several in Chapter 1 and will elaborate further on them in Chapter 7), then government should (again, a normative "should") intervene so as to rectify the inefficiency (assuming that the government agent knows how to do this).

Let's get on with our positive analysis. The start of predicting and understanding the behavior of a market is to comprehend the decisions made by a representative participant in the market. Chapter 2 is all about the interplay between what a consumer is able to do in the marketplace as depicted by a budget constraint, and what the consumer would like to do as described by a utility function. Working in a simple, two-good, frictionless environment, we analyzed three trade-off situations that involve, respectively: (1) two undefined, generic goods, *X* and *Y*, (2) two highly aggregated goods, current consumption and future consumption, and (3) two attributes of a financial portfolio, risk and expected return. In the current chapter we start with the two undefined goods, *X* and *Y*, and then return to our analysis of the risk and return characteristics of a financial portfolio.

We saw with regard to the *X*, *Y* trade-off, that the consumer's budget constraint in a two-good environment reflects three key economic variables: the price of *X*, the price of *Y*, and income. From the three, two in essence establish the constraint: the price of *X* *relative to* the price of *Y*, and nominal (dollar) income *relative to* the price of *Y* (or, we could have arbitrarily assessed nominal income relative to the price of *X*). The other important determinant of demand, the consumer's taste, is described by the decision maker's utility function and its related family of indifference curves.

Bringing these variables together, we showed the utility maximizing basket (the optimal *X*, *Y* basket; the optimal current consumption, future consumption basket; and the optimal risk, return portfolio configuration) for the decision maker to select. The optimal basket delivered equality between (1) the ratio of the marginal values assigned by the decision maker to each component of the basket (i.e., the decisions maker's marginal rate of substitution of $X$ for $Y$), and (2) the relative market prices of the two components of the basket (i.e., the price ratio, $P_X/P_Y$).

After establishing equilibrium for the consumer (given the two prices, his or her income, and his or her tastes), we moved on in Chapter 2 to comparative statics. To this end, we showed equilibrium for two different values of the price variable. We did this for the price of $X$ for the first trade-off situation, the price of current consumption for the second trade-off situation, and the price of risk for the third. This is where we pick up the story in the current chapter. We begin by considering the $X$, $Y$ trade-off and the household's demand curve for the generic, undefined good, $X$.

## THE DEMAND CURVE FOR *X*

The comparative statics diagram, Exhibit 2.12(b) in the previous chapter, presents all that we need in order to obtain an individual's demand curve for the generic, unidentified consumption good, $X$. The diagram is reproduced in Exhibit 3.1 with one more price, quantity combination added, and with a line drawn through its three illustrative points. That line is the consumer's demand curve for $X$. The letter "*d*" labels it as such.

In Exhibit 3.1, the quantity $X$ is the dependent variable and the price of $X$ is the independent variable. The usual mathematical convention is to show the

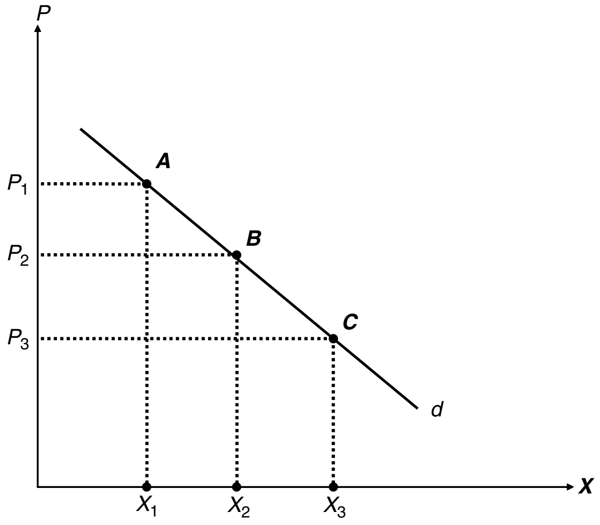


**EXHIBIT 3.1** A Consumer's Demand Curve for $X$

dependent variable on the vertical axis, and the independent variable on the horizontal axis. The axes are switched for the demand curve, however, and, if you think about it, this can be a bit confusing. If the price, quantity demanded relationship is described algebraically, we write a *demand function*,

$$X = f(P_X) \tag{3.1}$$

where $X$ is clearly the dependent variable. For simplicity, let's assume that the demand function is linear:

$$X = a - bP_X \tag{3.2}$$

The function rewritten with $P_X$ on the left-hand side is

$$P_X = \frac{a}{b} - \frac{1}{b}X \tag{3.3}$$

The intercept on the vertical axis is now seen to be $a/b$, and the slope of the curve is $-1/b$.

We suggest that you look at Exhibit 3.1, and at equations 3.2 and 3.3, and ask yourself the following question: All else equal, is a larger absolute value of $b$ associated with $X$ being more price-sensitive or less price-sensitive? The answer is found in equation 3.2: we see that $dX/dP_X = -b$. Thus a larger absolute value of $b$ means that the demand for $X$ is more responsive to the price of $X$. Turning to Exhibit 3.1, this means that the flatter is the demand curve for $X$ (i.e., if it is rotated out and up in a counterclockwise direction), the more the demand for $X$ responds to its price. Similarly, turning to equation 3.3, the larger is $b$, the less $P_X$ changes with $X$ (this is because $dP_X/dX = -1/b$).

This result can easily be verified. Go to Exhibit 3.1 and draw in a flatter curve that passes through point $B$ with the coordinates labeled $P_2$, $X_2$. You will see that with this flatter demand curve the quantity demanded at $P_1$ is less than $X_1$, and that the quantity demanded at $P_3$ is greater than $X_3$. In other words, your drawing should show explicitly that demand responds more to price when the curve is flatter.

Do not be misled, however. We do not mean to suggest that the slope measures responsiveness, only that a change in the slope of the linear curve changes responsiveness. Our *measure* of responsiveness remains elasticity.

There are a few more things to note about the representative consumer's demand curve:

1. The curve is downward sloping. The negative slope follows from the comparative statics presented in Chapter 2: the higher is $P_X$, the lower is the demand for $X$.

2. The demand curve shown in Exhibit 3.1 and the demand function described by equation 3.1 are linear. Our motive for sticking with linearity is simplicity only. Theory can predict slope, but generally does not take us any further than that. Empirical estimation must be relied on to gain additional insight into the shape of the downward sloping demand curve for any particular product being considered.
3. Other demand-determining variables (e.g., the price of the related good $Y$, income, and tastes) are constant along the demand curve. A change in any of these other independent variables will shift the demand curve.
4. If $X$ is a flow variable (e.g., the number of bottles of soda that are consumed *over a period of time*), then the quantity of $X$ is expressed *per unit of time* (e.g., per month or per year). If $X$ is a stock-dimensioned variable (i.e., the number of equity shares that are held), then $X$ is expressed as of a specific *point in time*.

## THE SENSITIVITY OF THE DEMAND FOR *X* TO THE PRICE OF *X*, THE PRICE OF *Y*, AND INCOME

Our analysis of micro markets places much attention on the price variable. In fact, microeconomics has at times been called *price theory*. How important is the price variable as a determinant of the demand for $X$? The price elasticity of demand answers this question. If the elasticity is high, price is important. If it is low, price does not matter very much. We introduced elasticity as a measure of responsiveness in Chapter 1. We continue with that discussion now.

But first, we should (a normative "should") be more specific. By "price elasticity of demand" we in fact mean, in the current context, *own price elasticity*. In other words, the elasticity of demand for $X$ with respect to $X$'s own price, $P_X$. The demand for $X$ might also be responsive to the price of another product, e.g., $Y$. The responsiveness of $X$ to $P_Y$ is a *cross-price elasticity*. The demand for $X$ is also responsive to income, $M$, and we further consider the *income elasticity of demand*. Here are the three elasticities spelled out. Note that each is the percentage change of $X$ divided by the percentage change of the independent variable:

$$\textbf{Own price elasticity:}\quad \eta_{X \cdot P_X} = \frac{dX}{X} \div \frac{dP_X}{P_X} = \frac{dX}{dP_X}\frac{P_X}{X} \tag{3.4}$$

$$\textbf{Cross-price elasticity:}\quad \eta_{X \cdot P_Y} = \frac{dX}{X} \div \frac{dP_Y}{P_Y} = \frac{dX}{dP_Y}\frac{P_Y}{X} \tag{3.5}$$

$$\textbf{Income elasticity:} \quad \eta_{X \cdot M} = \frac{dX}{X} \div \frac{dM}{M} = \frac{dX}{dM}\frac{M}{X} \tag{3.6}$$

## Income Elasticity

In our two goods formulation, income must be totally allocated between *X* and *Y*. Thus if income changes by any multiple, *k*, total expenditures (the total expenditure for *X* plus the total expenditure for *Y*) must change by the same multiple, *k*. If the demand for *X* increases by a multiple greater than *k* (that is, if the income elasticity of demand for *X* is greater than one), then the demand for *Y* must increase by a multiple less than *k* (that is, the income elasticity of demand for *Y* must be less than one). It is also possible for the income elasticity for one (but not both) of the two goods to be negative. This negative response can occur if, at the same time, the income elasticity of demand for the other good is sufficiently greater than one. The demand response to a change in income depends not only on the shape of the consumer's *X*, *Y* indifference curve, but also on the curve's location in *X*, *Y* space, for location as well as curvature determines where the indifference curve is tangent to the budget constraint.

Exhibit 3.2 displays a mapping where the demand for *X* rises with income at the same rate as the demand for *Y* (that is, where $\eta_{X \cdot M} = \eta_{Y \cdot M} = 1$). The figure contains two budget constraints ($P_X$ and $P_Y$ are constant, while income is assigned the two values, $M_1$ and $M_2$, respectively), and two indifference curves ($U_1$ and $U_2$). Tangency between the two budget constraints and the two indifference curves establishes the two optimal *X*, *Y* baskets labeled *A* and *B*. We have located A and B so that they lie on a straight line from the origin. The line that passes through the points of tangency is called an *income expansion path*. When the income expansion path is linear and intercepts the vertical axis at the origin, the income elasticity of demand is unity for each of the two goods (i.e., total expenditures for each of the two goods, $P_X X$ and $P_Y Y$, and income all increase at the same rate).

Two other possible tangency points labeled *C* and *D* along the $M_2$ budget line are also shown in Exhibit 3.2. If, for instance, the $U_2$ indifference curve were to lie further to the right along the $M_2$ budget constraint and be tangent to it at a point such as *C*, the income expansion path would be curving to the right, and the consumption of *X* would be increasing at a faster rate than the consumption of *Y*. In this case we would have $\eta_{X \cdot M} > 1 > \eta_{Y \cdot M}$. On the other hand, if the $U_2$ indifference curve were to lie further to the left along the $M_2$ budget constraint and be tangent to it at a point such as *D*, the income expansion path would be curving to the left, and the

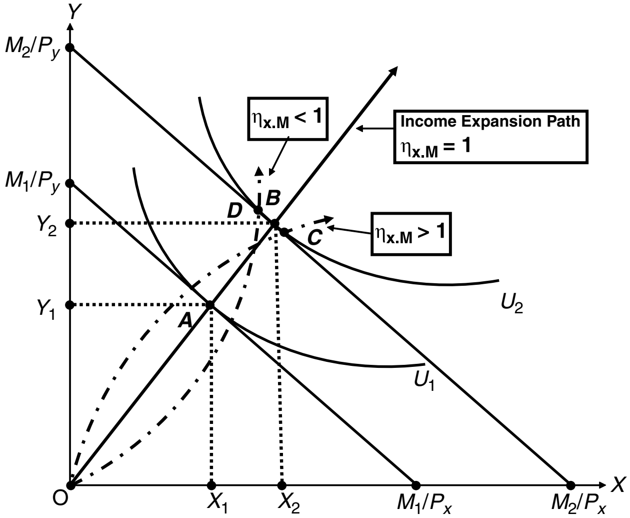


**EXHIBIT 3.2** Use of Indifference Curves and Budget Constraints to Determine an Individual's Income Expansion Path

consumption of *Y* would be increasing at a faster rate than the consumption of *X*. In this case we would have $\eta_{X \cdot M} < 1 < \eta_{Y \cdot M}$.

## Own Price Elasticity

Let's continue with our focus on good *X* and return to its own price elasticity of demand. The first task is to go back to the discussion in Chapter 1 and explain again why elasticity, not slope, is our measure of responsiveness. Now that we have established the demand curve, we can approach the explanation from a somewhat different angle by asking the following three questions.

1. What effect should changing the units in which the good *X* is measured have on our assessment of the responsiveness of the demand for *X* to its price?
2. What effect does the change have on the slope parameter of *X*'s demand curve?
3. What effect does the change have on the elasticity of demand for *X*?

Numerous examples exist of repackaging a product and, in so doing, changing the unit of measurement used for the quantity of a good sold. For instance, Ping-Pong balls could be sold separately or in packages of three.

The amount of cloth sold can be measured in feet, or in yards, or in centimeters. A stock selling at \$90 a share can have a three-for-one stock split. What do these measurement changes imply about price?

To understand the pure effect of the change, assume for simplicity that the repackaging is associated with nothing else (such as the offering of a quantity discount). This being the case, the price per three Ping-Pong balls is simply three times the price of one Ping-Pong ball. Or, the price per yard is three times the price per foot. Or, a stock trading at \$90 before a three-for-one split will trade at \$30 after the split. With these examples in mind, let's return to our generic product, *X*.

Assume that *X* is repackaged so that one new unit of *X* (call it $X'$) contains *t* times as many items as an old unit (call it *X*). This being the case, the number of units of $X'$ sold is the number of units of *X* sold divided by *t*. However, because one unit of $X'$ contains *t* units of *X*, the price of a unit of $X'$ is *t* times the price of a unit of *X*. In simple equations with *X* and $P_X$ on the left-hand side we have,

$$X = tX' \tag{3.7}$$

$$P_X = \frac{P'_X}{t} \tag{3.8}$$

Let the demand function for *X* be the linear equation,

$$X = a - bP_X \tag{3.9}$$

Accordingly, the elasticity of *X* with respect to $P_X$ is

$$\eta_{X\cdot P_X} = \frac{-bP_X}{X} \tag{3.10}$$

Substituting for *X* and $P_X$ in equation 3.9 we get

$$tX' = a - b\left(\frac{P'_X}{t}\right) \tag{3.11}$$

Thus,

$$X' = \frac{a}{t} - \left(\frac{b}{t^2}\right)P'_X \tag{3.12}$$

Now for the punch line: We changed the unit of measurement for *X*, making *t* old items equal to one new unit, and the slope of the linear demand function, $dX'/dP'_X$ is now the slope of the old demand curve ($dX/dP_X = -b$) divided by $t^2$. Does this mean that the demand for the good has suddenly and mysteriously become more price-insensitive after the product has been repackaged?

We can now answer the three questions raised previously:

**Question 1:** "What effect should changing the units in which the good *X* is measured have on our assessment of the responsiveness of the demand for *X* to its price?" Answer: It should have no effect.

**Question 2:** What effect does the repackaging have on the slope parameter of *X*'s demand curve? We have already answered the second question: The slope is divided by $t^2$. But we just said that our measure of responsiveness should not be affected. Apparently, the slope parameter is not a very good measure of responsiveness.

**Question 3:** What effect does the repackaging have on the elasticity of demand for *X*? To answer this one, write the elasticity of demand of $X'$ with respect to $P'_X$ as

$$\eta_{X' \cdot P'_X} = \left(\frac{dX'}{dP'_X}\right)\left(\frac{P'_X}{X'}\right) \tag{3.13}$$

From equation 3.12 we have $dX'/dP'_X = -b/t^2$. Using this along with $P'_X = tP_X$ (from equation 3.8) and $X' = X/t$ (from equation 3.7), we get

$$\eta_{X' \cdot P'_X} = \left(-\frac{b}{t^2}\right)\left(\frac{tP_X}{X/t}\right) = (-b)\left(\frac{P_X}{X}\right) = \eta_{X \cdot P_X} \tag{3.14}$$

And so we have it, elasticity is not affected by any change in the unit of measurement that is due to repackaging.

Once again, elasticity wins as our measure of responsiveness.

## Own and Cross-Price Elasticity

Now, on to the next question: Why is the demand for some goods more elastic than the demand for other goods? Good *X*'s own price elasticity of demand (and cross-price elasticity of demand) reflects the ease with which the consumer can substitute between good *X* and good *Y*. Goods that can be used in place of one another are *substitutes*. Goods that are best used in combination with one another are *complements*. Unless goods are perfect complements, however, some substitutability is still possible. For our representative consumer, butter and margarine, for instance, are very close substitutes, while beer and pretzels in the evening hours might better be viewed as complements. Even better, left shoes and right shoes are best described as being perfect complements.

If $X$ and $Y$ are relatively good substitutes for each other (such as the previous butter and margarine example), the $X$, $Y$ indifference curves are only slightly curved (the $MRS_{X.Y}$ does not change much as we move along one of the curves). But if $X$ and $Y$ are good complements (such as tennis rackets and tennis balls), the absolute value of the slope of the $X$, $Y$ indifference curve changes sharply from a large number to a small number in the neighborhood where $X$ and $Y$ would naturally be combined. In the extreme case of perfect complementarity, the $X$, $Y$ indifference curves are L-shaped.

Except for an extreme left shoe, right shoe situation where two goods are perfect complements, a two-good universe pretty much forces the two goods to appear to be substitutes. That is, when income can be allocated between two goods only, it is difficult for the consumption of both of the goods to increase following a price decrease for one of them (e.g., for the goods to be complements) although this can happen because of the income effect that attends a price decrease. In any event, complements can be defined more clearly in a multigood environment:

1. $X$ and $Y$ are substitutes if the consumption of *more* $Y$ is associated with the consumption of *less* $X$.
2. $X$ and $Y$ are complements if the consumption of *more* $Y$ is associated with the consumption of *more* $X$.

We can go one step further and classify two goods as substitutes and complements based on the sign of their cross elasticity:

1. $X$ and $Y$ are substitutes if the cross elasticity of demand for $X$ with respect to the price of $Y$ is *positive*.
2. $X$ and $Y$ are complements if the cross elasticity of demand for $X$ with respect to the price of $Y$ is *negative*.

Here is another way to look at it:

1. With substitutes, the own price elasticity and the cross-price elasticity have opposite signs (own is negative and cross is positive).
2. With complements, the own elasticity and the cross elasticity have the same sign (they are both negative).

We display two alternative indifference curves in Exhibit 3.3. The top diagram [Exhibit 3.3(a)] presents one indifference curve (labeled $U_1$) and

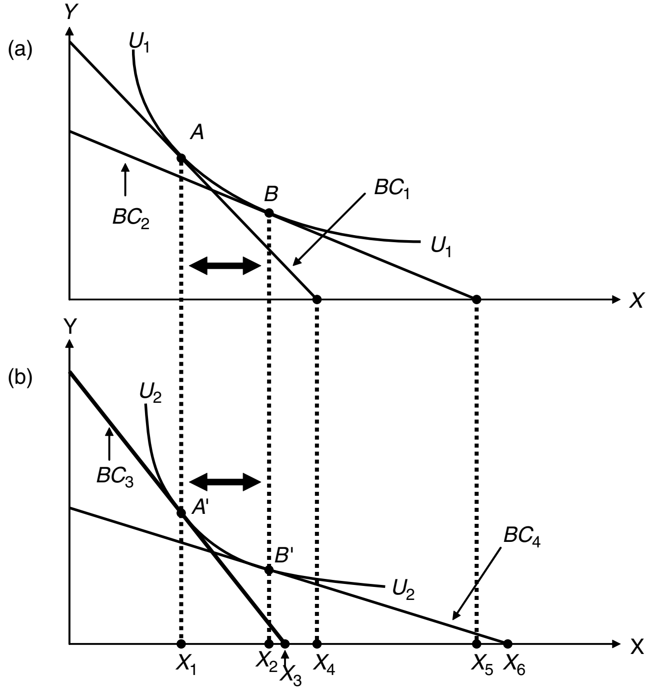


**EXHIBIT 3.3** Demonstration That the Curvature of an Indifference Curve Reflects the Extent to Which *X* and *Y* Are Relatively Good Substitutes [3.3(a)] or Complements [3.3(b)]

two budget constraints (labeled $BC_1$ and $BC_2$). The bottom diagram [Exhibit 3.3(b)] also presents one indifference curve (labeled $U_2$) and two budget constraints (labeled $BC_3$ and $BC_4$). Look at the two indifference curves, $U_1$ and $U_2$. There is no presumption that one of these indifference curves represents a higher level of utility than the other; our focus here is only on substitutability. Which curve in your opinion shows that *X* and *Y* are better substitutes, $U_1$ or $U_2$, and which shows that they are better complements?

Have you paused a moment and thought about it? Can you see that the $U_2$ curve is more curved than the $U_1$ curve? This indicates that *X* and *Y* are relatively good substitutes in the top diagram, and that they are relatively poorer substitutes in the lower diagram.

Discerning the difference in curvature by eye is not that easy. Exhibit 3.3 gives us another way of understanding that *X* and *Y* are portrayed as being better substitutes in the top diagram than in the bottom diagram. In Exhibit 3.3(a), the $BC_1$ constraint is tangent to the $U_1$ indifference curve at the basket labeled A, and the $BC_2$ constraint is tangent to the same $U_1$

indifference curve at the basket labeled B. The difference in the slope between the $BC_1$ constraint and the $BC_2$ constraint reflects the magnitude of the change in $P_X$ that is required to get the consumer to substitute $X$ for $Y$ by changing from basket A to basket B, a change that increases the consumption of $X$ from $X_1$ to $X_2$.

Vertical lines have been dropped from points $A$ and $B$ in the upper diagram to the $X$ axis of the lower diagram. These lines enable us to identify baskets $A'$ and $B'$ in the lower diagram that contain the same amount of $X$ as, respectively, baskets $A$ and $B$ in the top diagram. What slope must the budget line have for basket $A'$ to be selected in Exhibit 3.3(b)? The budget line must have the slope represented by the steeper constraint $BC_3$ in the lower diagram. Note that $BC_3$ is steeper than $BC_1$, its counterpart in Exhibit 3.3(a).

What slope must the budget line have for basket $B'$ to be selected in the lower diagram? It must have the slope represented by the flatter constraint $BC_4$ in Exhibit 3.3(b). Note that $BC_4$ is flatter than $BC_2$, its counterpart in the top diagram.

Let's put this all together. The exhibit shows that the price of $X$ has to fall more in the bottom diagram than in the top diagram for the consumer to increase his or her consumption of $X$ from $X_1$ to $X_2$. This is because $X$ and $Y$ are better substitutes in the top diagram than in the bottom diagram. Does this make intuitive sense to you? Try thinking about it this way: It is easier to induce a switch of demand from one good to the other when the two goods are better substitutes, which means that a smaller price change is required to induce a switch from $A$ to $B$ than from $A'$ to $B'$.

Perhaps your eye cannot easily pick up that $BC_3$ is steeper than $BC_1$, its counterpart in the top diagram, or that $BC_4$ is flatter than $BC_2$, its counterpart in the top diagram. As a further visual aid, we have identified four additional quantities on the $X$ axis of the lower diagram: $X_3 < X_4 < X_5 < X_6$. These are the $X$ axis intercepts of, respectively, $BC_3$ in lower diagram (the highest price of $X$ for the four constraints), $BC_1$ in the upper diagram (the second highest price of $X$), $BC_2$ also in the upper diagram (the second lowest price of $X$) and finally $BC_4$ in the lower diagram (the lowest price of $X$ for the four constraints). These intercepts show that to get the consumer to demand $X_2 - X_1$ more units of $X$, the budget constraint had to be rotated considerably more in the lower diagram than in the top diagram. The $X$ intercept in the top diagram increased a relatively modest amount, from $X_4$ to $X_5$, while the $X$ intercept in the bottom diagram increased by the larger amount, from $X_3$ to $X_6$.

Now let's go to the extremes: *perfect substitutes* (the decision maker is totally indifferent between consuming $X$ and consuming $Y$) and *perfect complements* (additional $X$ is totally useless to the decision maker unless it

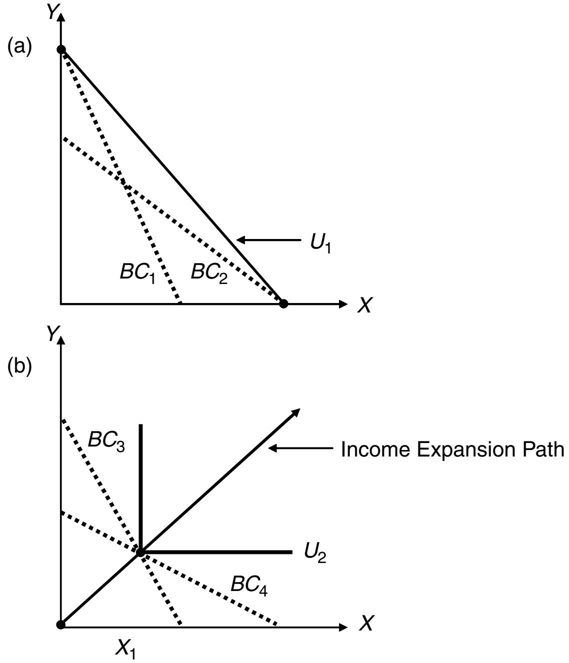


**EXHIBIT 3.4** Indifference Curves for Perfect Substitutes [3.4 (a)] and Perfect Complements [3.4 (b)]

is obtained in fixed proportions with additional $Y$, and vice versa). The two diagrams in Exhibit 3.4 show representative indifference curves for perfect substitutes (the curve labeled $U_1$ in the top diagram) and for perfect complements (the curve labeled $U_2$ in the bottom diagram).

In Exhibit 3.4, the $U_1$ indifference curve for perfect substitutes is linear. The constant slope of the curve shows the constant rate at which $X$ can be substituted for $Y$. If the units of measurement for the two goods are the same and substitution is a one-for-one proposition, then the slope of the curve (the negative of $MRS_{X \cdot Y}$) is $-1$. For our purposes, we need not concern ourselves with the substitution ratio, for all that matters is that the $MRS_{X \cdot Y}$ is constant. This being the case, the decision maker will, in general, go to a corner solution (i.e., consume only $X$ or only $Y$ depending on the relative prices of the two goods).

Two budget constraints are displayed in Exhibit 3.4(a). For the constraint labeled $BC_1$, we have $P_X/P_Y > MRS_{X \cdot Y}$, and the decision maker buys only $Y$. For the constraint labeled $BC_2$, we have $P_X/P_Y < MRS_{X \cdot Y}$, and the decision maker buys only $X$. For the special case where $P_X/P_Y = MRS_{X \cdot Y}$, the budget constraint is superimposed on the indifference curve, and the decision maker could arbitrarily select his or her *X*, *Y* basket anywhere along the constraint. It follows that the demand curve for *X* (and also for its perfect substitute, *Y*) is a horizontal line with an intercept on the vertical axis at $P_X = MRS_{X \cdot Y} P_Y$. The horizontal curve is infinitely elastic (because for a slight increase in $P_X$, $X = 0$ and thus $P_X/X =$ infinity).

In Exhibit 3.4(b), the $U_2$ indifference curve for perfect complements is shown to be L-shaped. A linear income expansion path rises from the origin and passes through the corner of the L-shaped curve. For perfect complements in a two-good environment, the income expansion path is always a positively inclined straight line from the origin, and it passes through the corner of any L-shaped indifference curve that we might draw. The slope of the income expansion path is determined by the number of units of *Y* that must be used in combination with one unit of *X*, with a one-to-one combination translating into a slope of +1. For our purposes, we need not concern ourselves with what the *X*/*Y* ratio is, for all that matters is that:

1. To the left of the income expansion path the indifference curve is vertical—the marginal utility of *Y* is zero (more *Y* is useless unless it has more *X* to go with it), and thus the $MRS_{X \cdot Y}$ is infinite.
2. To the right of the income expansion path the indifference curve is horizontal—the marginal utility of *X* is zero (more *X* is useless unless it has more *Y* to go with it), and thus the $MRS_{X \cdot Y}$ is zero.

Two budget constraints, $BC_3$ and $BC_4$, are displayed in Exhibit 3.4(b). Both constraints touch the L-shaped indifference curve at its corner and are below it otherwise. For any move away from the corner in the direction of more *Y*, we would have $P_X/P_Y < MRS_{X \cdot Y}$, and for any move away from the corner in the direction of more *X*, we would have $P_X/P_Y > MRS_{X \cdot Y}$. Accordingly, the optimal basket to select lies at the corner of an L-shaped curve. This is hardly surprising: after all, *X* and *Y* have to be used in a fixed combination.

What this analysis highlights is that the optimal consumption basket, in terms of proportions, is invariant with regard to the price of *X* relative to the price of *Y*. If the price of either of these two goods were to change (e.g., if the price of *X* were to decrease, with $P_Y$ and income constant), demand for each of the two goods would change (e.g., more *X* and more *Y* would be demanded), but the response would be attributable entirely to the income effect—the substitution effect would be zero. If in an *N*-good world, *X* is a small budget item, the income effect is practically zero, and the own price elasticity would be nearly zero (i.e., the demand curve would be close to vertical). In the *N*-good world, it may be preferable to consider how an *X*, *Y* bundle, as a bundle, would face substitutes in other goods (or bundles of goods).

How easy is it, you might ask, to classify goods as substitutes or as complements simply by looking at how they are used? Classifying two goods in this way might be straightforward in some cases (left shoes and right shoes are obvious complements, and butter and margarine quite clearly are pretty good substitutes). In other cases, the classification may not be as obvious. Consider, for instance, eggs and flour. Various cake recipes call for both of these ingredients, but not necessarily in the same proportions. Assume that cake A is relatively egg-intensive, and that cake B is relatively flour-intensive. Recognizing this, ask yourself the following question: Are eggs and flour complements or substitutes?

The baking of each cake follows a recipe that calls for the two inputs to be used in strict, "technologically given" proportions. Hence, for each cake, eggs and flour are perfect complements. But between the cakes, flour and eggs are substitutes. That is, flour can be substituted for eggs by substituting some of cake B (the flour-intensive cake) for some of cake A (the egg-intensive cake). This substitution should (a positive "should") occur if the price of eggs were to rise relative to the price of flour. People would consume fewer A cakes and would dine more on B cakes.

### A Concluding Remark

Following our eggs and flour example, we conclude this discussion of the sensitivity of *X* to the price of *X* with the following thought. Substitutability drives markets: It is at the heart of microeconomic analysis. Substitutability means that choice exists, that there are alternative ways of reaching our goals, that consumption (and also production) decisions are not simply dictated by technological coefficients, as they would be in an environment where all action is governed by formulas (which is what recipes are), and where perfect complements therefore prevail. Looking for the substitution possibilities, understanding them, and exploiting them is the challenge that decision makers face. And it is the responses of the decision makers when they are confronted with meaningful choices that make our micro markets dynamic.

## FROM INDIVIDUAL DEMAND CURVES TO MARKET DEMAND CURVES

We have now established an individual's demand curve to consume the generic product, *X*. We have examined its properties, paying particular attention to *X*'s elasticity of demand with respect to three of our basic

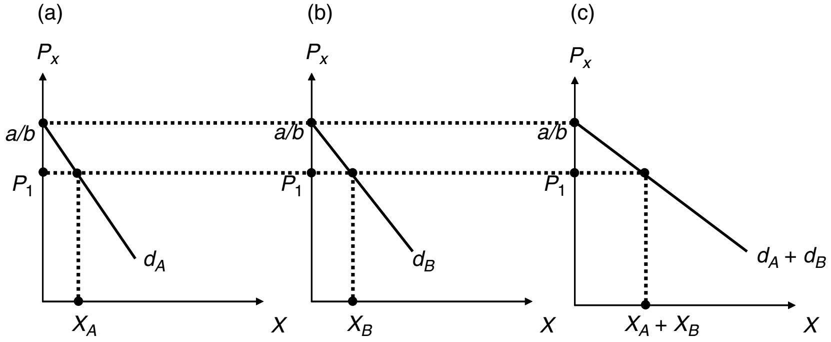


**EXHIBIT 3.5** The Aggregation from Individual Demand Curves [3.5(a) and (b)] to the Market Demand Curve [3.5(c)]

demand-determining variables: X's own price ($P_X$), the price of the other good ($P_Y$), and the decision maker's income (M). Throughout, we have taken tastes as given and constant (i.e., tastes have received the *ceteris paribus* treatment). Now we progress from the demand curve of an individual to the demand curve of the market.

By market demand, we are referring to the aggregate of the demands of all of the customers in the market. After obtaining the market's aggregate demand curve, we can match it with the market's supply curve for the good and see how price and quantity are both established in a frictionless marketplace that comprises a large number of decision makers, none of whom have any power individually over the market price.

Aggregation from individual demands to the market demand is a simple process. To see how it works, take a look at Exhibit 3.5. Three diagrams comprise the exhibit. Exhibit 3.5(a) depicts the first individual's demand curve ($d_A$), the second diagram, 3.5(b), depicts the second individual's demand curve ($d_B$), and the third diagram, 3.5(c), depicts the demand curve that is an aggregate of these two individual curves ($d_A + d_B$). For simplicity, let the two individuals have identical demand curves.

As you can see in the diagram, aggregation consists of horizontally summing $d_A$ and $d_B$ at each of the prices. Horizontal summation means that, at each price, the quantities shown on the horizontal axis are added together (vertical summation would mean that, at each quantity, the prices shown on the vertical axis are added together).

The two individual demand curves each have the same intercept ($a/b$) on the vertical axis and the same slope, $dP/dX = -1/b$. Exhibit 3.5(a) shows that at the price $P_1$, the first individual demands $X_A$ units; 3.5(b) shows that the second individual demands $X_B$ units (which in the current case equals

$X_A$), and 3.5(c) shows that the two individuals together demand $X_A + X_B$ (which equals $2X_A$).

Algebraically stated, following equation 3.2, the first consumer's demand curve is

$$X_A = a - bP \tag{3.15}$$

Setting $X_A = 0$, we see that the intercept on the vertical axis is $P = a/b$, as it is in equation 3.3. Because we have (for simplicity only) assumed that the second consumer has an identical demand curve, the demand for both consumers together is

$$X_{A+B} = X_A + X_B = 2(a - bP) \tag{3.16}$$

Setting $X_{A+B} = 0$, we again see that the intercept on the vertical axis remains fixed at $P = a/b$. Equation 3.16 should further clarify what is meant by saying that, to get the aggregated demand curve for the two consumers, we sum the individual curves horizontally (i.e., add the quantities at each price).

The slope of the aggregated curve is $-2b$, which is consistent with the demand curve in Exhibit 3.5(c), rotating up and to the right to a position where $dP/dX_{A+B} = (^1\!/\!_2)(dP/dX)$. If you find this tedious and confusing, blame it on the convention of putting the independent variable (price) on the vertical axis. But, it helps to be patient. You will understand why this is done when we get to Chapter 5.

Now assume that not two, but $N$ identical (again, for simplicity only) consumers (indexed $1, \ldots, N$) comprise the market for $X$. Equation (3.16) simply becomes

$$X^D = X_1 + \ldots + X_N = N(a - bP) \tag{3.17}$$

where $X^D$ is the aggregated market demand for $X$. Equation 3.17 is what we want; it is the market's (linear) demand function for $X$.

## Elasticity Revisited

Having obtained the demand curve for the aggregate of $N$ (identical) consumers, we can see again why we use elasticity rather than slope as the measure of responsiveness. The demonstration is equivalent to the one that we presented above to identify the effect of repackaging on responsiveness.

From equation 3.17 we have $dX^D/dP = N(dX^{IND}/dP) = -Nb$, where $X^{IND}$ is an individual's demand. Accordingly, if we took the slope parameter as the measure of responsiveness, we would say that the market's aggregate demand was $N$ times as responsive as each of the individual demands, even though the individual demands are all identical to each other. This does not make sense, this cannot be. Accordingly, we again turn to elasticity instead of slope as our measure. Recognizing that, at a given price, $X^D = NX^{IND}$, our elasticity measure for the market ($\eta_{X\cdot P}^{D(MKT)}$) is

$$\eta_{X\cdot P}^{D(MKT)} = -Nb\left(\frac{P}{NX^{IND}}\right) = -b\left(\frac{P}{X^{IND}}\right) = \eta_{X\cdot P}^{D(IND)} \tag{3.18}$$

We trust that, by now, we have convinced you that elasticity, not slope, is the appropriate measure of responsiveness. We have, however, one final point to make regarding elasticity: it is not constant along a linear demand curve. If price is high and quantity small, a small percentage decrease in price will induce a larger percentage increase in quantity than if quantity is large and price is low. In other words, as we move down along the negatively inclined curve, demand becomes increasingly unresponsive to further price decreases. Perhaps you have already thought of this. After all, we have been writing elasticity as

$$\eta_{X\cdot P} = \frac{dX}{dP}\frac{P}{X} = -b\frac{P}{X} \tag{3.19}$$

It is clear from equation 3.19 that, with the slope parameter, $b$, constant, $\eta_{X\cdot P}$ decreases in absolute value as $P$ falls and $X$, in response, rises, as we move down along the linear demand curve. Be sure to understand, however, that a changing price elasticity is not necessarily a property of a demand curve, although it does characterize a linear demand curve.

This completes our discussion of the demand curve for the undefined, generic good, $X$.

## THE DEMAND TO HOLD SHARES OF A RISKY ASSET

In that section we consider demand analysis in a more defined context: the demand to hold shares of a risky asset. In so doing, we build on material presented in Chapter 2.

In our earlier discussion, we first considered the risk, return trade-off as represented by the investor's family of indifference curves (the depiction of his or her tastes) and by the capital market line (the depiction of possibilities in the marketplace). Tangency between the capital market line and an indifference curve established the optimal risk and return configuration for the investor. We showed how the optimal configuration changes as the capital market line shifts. Then, we made share price and the number of shares held explicit variables; this translation brought us to the investor's demand curve to hold shares. We saw that, holding the risk-free rate of interest ($r_f$) constant, the expected future share price [$E(P_T)$] constant, and tastes for risk constant, the investor will want to hold more shares of the risky portfolio at a lower current price per share.

For our analysis in Chapter 2, we assumed frictionless trading and a two asset world, which includes a risky asset with an expected return of $E(r_m)$, and a riskless asset that yields a return of $r_f$ with certainty. We continue to operate in this environment now. As in Chapter 2, you might think of the single risky asset as being a very broadly based index fund or an exchange-traded fund. Starting in this context with a utility of wealth function and a set of additional simplifying assumptions (we always do this simplification thing), it is possible to derive more formally the investor's downward sloping, linear demand curve to hold shares of the risky asset.[1]

## THE INVESTOR'S DEMAND CURVE FOR THE RISKY ASSET

A representative investor's linear demand to hold shares of the risky market portfolio is shown in Exhibit 3.6. The first thing to note about the curve is the definition of the dependent variable $N$, *the number of shares held*. This is a stock (no pun intended) measure, not a flow-dimensioned measure. The "stock"-dimensioned measure, "$N$ shares of stock," is the number of shares held as of a *specific point in time*. These shares are assets that are being held by the investor. As with all assets, the accounting for the equity shares is, as we have just said, as of a specific point in time, and hence it is a stock dimensioned variable.

The second thing to note about the curve is its price intercept, $E(P_T)/(1 + r_f)$. The intercept term is the price at which the investor would hold precisely zero shares of the risky asset (i.e., be entirely in the risk-free asset, the instrument that yields the return $r_f$ with certainty). Why would the investor want to be entirely in the risk-free asset at this price? If $P = E(P_T)/(1 + r_f)$, then $E(P_T) = P(1 + r_f)$. That is, if the investor paid the price represented by the intercept and if he or she expected price to be $E(P_T)$ at the end of his or her holding period, the investor, on expectation, is

---

[1] A proof is given in Thomas Ho, Robert A. Schwartz, and David K. Whitcomb, "The Trading Decision and Market Clearing under Transaction Price Uncertainty," *Journal of Finance* (March 1985), 21–42.

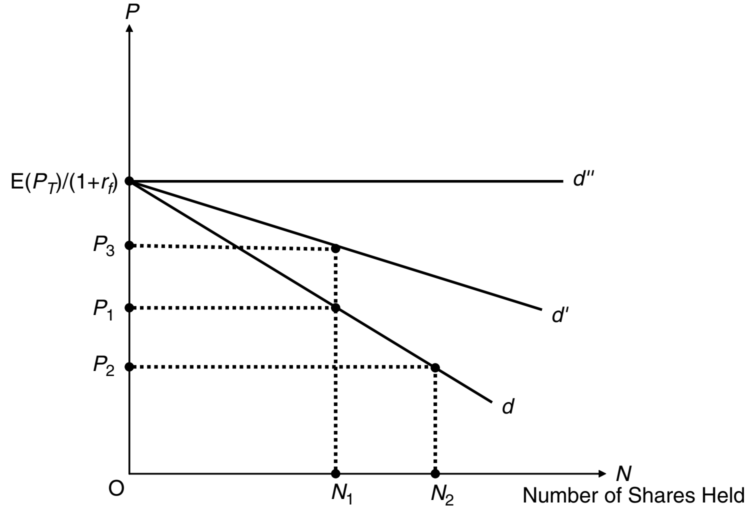


**EXHIBIT 3.6** An Investor's Demand to Hold Shares of a Risky Portfolio

realizing only the risk-free rate of return. The risk premium would be zero. But our discussion in Chapter 2 establishes that a risk-averse decision maker will require a risk premium before willingly accepting an avoidable risk, and, further, we assume that our representative investor is, indeed, risk-averse. So, a zero risk premium for a risky asset means that the optimal holding of the asset for a risk-averse investor is a big fat zero. That is precisely what the price intercept in Exhibit 3.6 is telling us.

The third thing to notice about the demand curve in Exhibit 3.6 is that it is downward sloping. This is because the investor is risk-averse. With the curve downward sloping, the investor will hold $N_1$ shares at the lower price of $P_1$, and a yet larger number of shares, $N_2$, at the yet lower price, $P_2$. At $P_1$ the risk premium is $\frac{E(P_T)}{P_1} - 1 - r_f$, while at the lower price of $P_2$, the risk premium is the larger amount, $\frac{E(P_T)}{P_2} - 1 - r_f$. So, the downward sloping demand curve tells us that, the larger the risk premium, the more shares of the risky asset our representative investor will want to hold.

A fourth thing can be noted about the demand curve in Exhibit 3.6. We have included a second curve labeled *d*′, which would describe a different taste for risk. Note that the two curves have the same intercept on the price axis, but that the second curve is not decreasing as rapidly. As an exercise in comparative statics, what can we say if *d*′ was the demand curve of a second investor? We would say that the second investor with the *d*′ demand curve is less risk-averse than the first investor whose demand curve is *d*. You can nail this thought down by identifying the price on the *d*′ curve at which the investor would want to hold $N_1$ shares. At $N_1$ shares, the *d′* price is $P_3$ (which is higher than the d price of $P_1$), and thus the risk premium must be less. If a lower risk premium would be acceptable for this second investor, than he or she must be less risk-averse than the first investor.

Now keep rotating the demand curve up until it is horizontal and thus infinitely elastic. We have labeled the infinitely elastic curve *d″*. What is the attitude toward risk of a third investor with the *d″* curve? What risk premium would this investor require to hold shares of the risky asset? The required risk premium would be zero and the third investor would have a risk-neutral utility function (namely, his or her utility as a function of wealth would be linear). For a third investor who is risk-neutral, the risky asset and the risk-free asset are perfect substitutes. He or she would be indifferent between them if the expected return on the risky asset equaled the certain return for the riskless asset. We saw earlier in this chapter that if two products (*X* and *Y*) are perfect substitutes, the demand curve for product *X* (and for *Y*) is infinitely elastic at a price that satisfies $P_X/P_Y = MRS_{X \cdot Y}$. For the risky asset, this price is $P = E(P_T)/(1 + r_f)$. At this price, the curve *d″* is infinitely elastic. However, we have assumed that all investors are risk-averse. For us, the third investor does not exist.

## EQUILIBRIUM IN THE MARKET FOR *X*

We have established the demand curve and its important properties for the undefined, generic good, *X*, and for the risky asset. Our next objective is to show how price is set in the marketplace for *X* (our topic in this section) and for the risky asset (our topic in the section that follows).

To obtain equilibrium values for $P_X$ and *X*, we need a supply curve to match with the demand curve. We can introduce the supply curve now. In so doing, we will point out the basic properties of the curve but will not derive it as we did the demand curve. In brief, just as utility functions and indifference curves underlie a demand curve, so do production functions and cost curves underlie a supply curve. We will show you how this works when we get to Chapter 5.

### The Supply Curve of X

A supply curve has the same basic properties as a demand curve. We show a supply curve for the market in Exhibit 3.7. The curve depicts the relationship between the amount of *X* that will be supplied and the price of *X*.

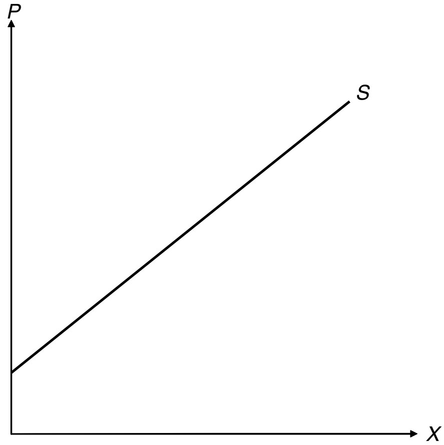


**EXHIBIT 3.7** The Market Supply Curve for *X*

Here are the important properties of the supply curve:

- As we explained with regard to demand curves, with a supply curve *X* is expressed as the number of units supplied per unit of time if *X* is a flow-dimensioned variable, and is expressed as the number of units that are available to be held if *X* is a stock-dimensioned variable.
- Unlike demand curves that are downward sloping, supply curves are generally upward sloping (although there are exceptions).
- Like the demand curve, the supply curve shows the relationship between the quantity of *X* and the price of *X*, all other relevant variables held constant (the *ceteris paribus* treatment). For the supply curve, the other variables include the price of other goods that are substitutes in production, the cost of inputs into the production of *X*, and technology. If one of the *ceteris paribus* variables changes, the supply curve shifts.
- Like a market demand curve, a market supply curve is the horizontal summation of the individual firm supply curves. That is, we again sum all of the quantities that will be brought to market by each of the individual firms at each possible value of price.
- As with a demand curve, the responsiveness of *X* supplied to the price of *X* is measured by

$$\textbf{Supply elasticity:}\ \eta_{X \cdot P_X} = \frac{dX}{X} \div \frac{dP_X}{P_X} = \frac{dX}{dP_X}\frac{P_X}{X} > 0 \qquad (3.20)$$

Note that the supply elasticity is generally positive while the demand elasticity is generally negative.

### The Intersection of Demand and Supply

We bring the supply curve and the demand curve together in Exhibit 3.8. The two curves intersect each other at the price $P^*$ and the quantity $X^*$. These two values are the equilibrium values of price and quantity on the market. They are stable equilibrium values. They are the values of price and quantity that the market will push price and quantity to. Exhibit 3.8(a) shows that, at any price higher than $P^*$, say $P_1$, an amount $X_1^S > X^*$ will be brought to market, and an amount $X_1^D < X^*$ will be demanded. What happens when $X_1^S > X_1^D$? Suppliers in aggregate will be scrambling to sell more units of $X$ to buyers than buyers, in aggregate, wish to purchase at the higher price. This competitive response on the part of the suppliers puts downward pressure on price, and the price of $X$ will fall. How far? To $P^*$, where the balance between selling pressure and buying pressure is restored.

Exhibit 3.8(b) shows that, at any price lower than $P^*$, say $P_2$, an amount $X_2^D > X^*$ will be sought for purchase, and an amount $X_2^S < X^*$ will be supplied. What happens when $X_2^D > X_2^S$? Buyers, in aggregate, will be scrambling to purchase more units of $X$ than sellers, in aggregate, will be bringing to market at the lower price. This competitive response on the part

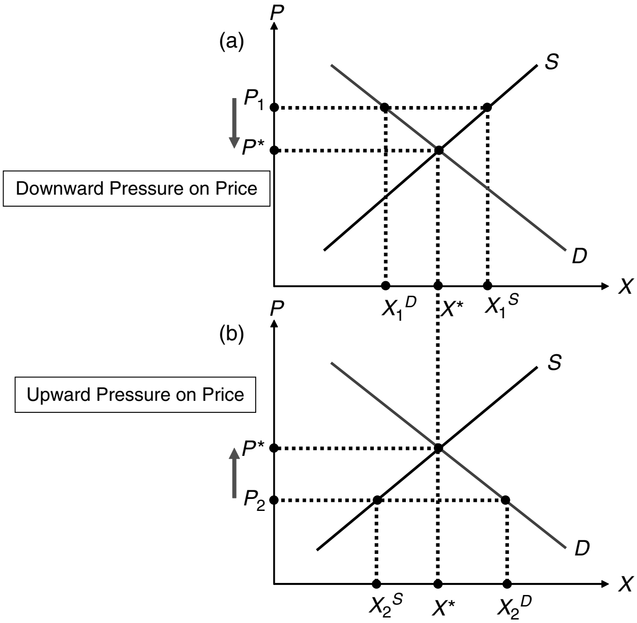


**EXHIBIT 3.8** The Intersection of the Market Demand Curve ($D$) and Supply Curve ($S$) Determine the Equilibrium Price ($P^*$) and Quantity ($X^*$). Downward Pressure on Price [3.8(a)], and Upward Pressure on Price [3.8(b)].

of the buyers puts upward pressure on price, and the price of $X$ will rise. How far? To $P^*$, where the balance between selling pressure and buying pressure is restored.

So, from both directions (that of excess supply and that of excess demand), economic forces will push price to $P^*$ whenever it is above or below its equilibrium value. And as price settles down at $P^*$, quantity will have found its equilibrium value of $X^*$. And this is a *stable value*. A market that has achieved this equilibrium has neither excess demand nor an overhang of supply. Its economic efficiency could not be better. Price will have done exactly what it should do as a rationing variable, and Adam Smith's invisible hand will deserve a solid round of applause.

Here is another way of looking at the diagrams in Exhibit 3.8. Interpret the demand curve as showing the maximum that will be demanded at each value of price, and interpret the supply curve as showing the maximum that will be supplied at each value of price. Ignoring unpalatable acts such as forced feeding, shot gun marriages, and other forms of conscription, the amount of $X$ that can change hands in the market is the minimum of these two maximums. At prices below $P^*$, the quantity supplied is the smaller of the two maximums, and in this range the quantity that can change hands is determined by the supply curve alone (and all of these quantities are smaller than $Q^*$). At prices above $P^*$, the quantity demanded is the smaller of the two maximums, and in this range the quantities are determined by the demand curve alone, (and all of the quantities that can change hands are also smaller than $Q^*$). At $P^*$ where the two maximums are equal, the largest transaction possible from both a demand and a supply consideration (namely, $X^*$) can be achieved. Conclusion: the free-market solution maximizes trading, and trading is what markets are all about!

## Regulatory Intervention in the Market Process

Let's take a closer look at what can happen if barriers that are placed on price prevent the $P^*$, $X^*$ equilibrium solution from being attained. In Exhibit 3.9 we present two diagrams, both of which display the same demand and supply curves and the same equilibrium values, $P^*$ and $X^*$ that are shown in Exhibit 3.8. In Exhibit 3.9(a), a price floor (labeled $P^{MIN}$) has been imposed at a price above $P^*$, and in Exhibit 3.9(b) a price ceiling (labeled $P^{MAX}$) has been imposed at a price below $P^*$. You might think of these two diagrams as representing the "floor above and the ceiling below."

As shown in Exhibit 3.9(a), the "floor above" situation results in a supply, demand relation where $X^D < X^* < X^S$. The free adjustment of price as a rationing variable has been blocked, the demand curve alone determines volume and excess supply results. How does the excess supply get rationed?

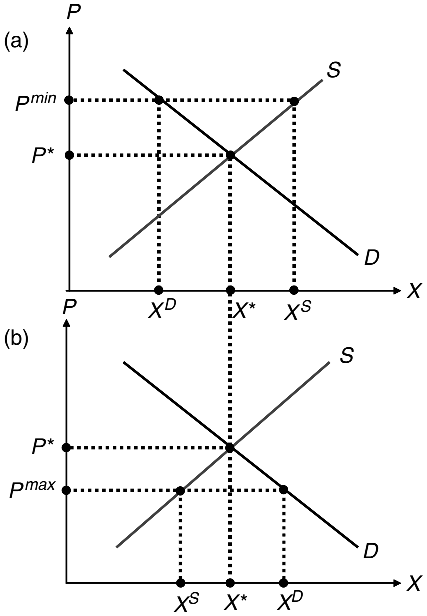


**EXHIBIT 3.9** The Rationing Problem Created by the Imposition of a Price Floor [3.9(a)] or a Price Ceiling [3.9(b)]

By the imposition of production quotas? Is the allocation determined by a first-come, first-served rule? By something else?

As shown in Exhibit 3.9(b), the "ceiling below" situation results in a supply, demand relation where $X^S < X^* < X^D$. Again, the free adjustment of price as a rationing variable has been blocked. This time the supply curve alone determines sales volume, and excess demand results. Who now gets the scarce supply? Is the allocation determined by a first come, first served rule? By pro rata rationing with everybody getting only a fraction of what they would like to consume at the price of $P^{MAX}$? By something else?

How has a floor above or a ceiling below been imposed? Here are five classic examples of imposed price controls: rent control (a ceiling), agricultural price supports (a floor), minimum wage laws (a floor), minimum trading commissions (a floor), and currency exchange rate limits (a floor and a ceiling). Each has been motivated by a desire to improve the economic welfare of the target (renters, farmers, unskilled labor, stock exchanges members, and the stability of a financial system). Each has created a rationing problem. Unfortunately, the rationing devices that have been resorted to in the absence of price acting freely as a rationing variable have typically had undesirable, unintended consequences.

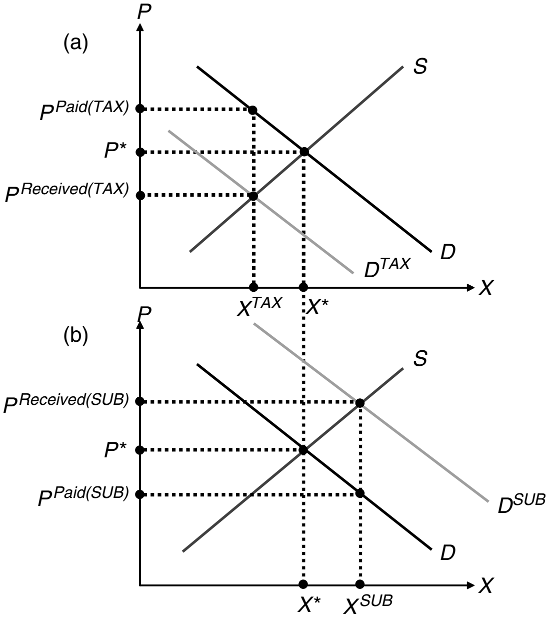


**EXHIBIT 3.10** The Effect of a Sales Tax [3.10(a)] and Subsidy [3.10(b)] on Price and Quantity

Government has further ways to intervene in the workings of a free market. One is the imposition of a sales tax or subsidy. We consider the effect of this intervention with reference to Exhibit 3.10. Exhibit 3.10(a) shows the effects of a tax, and Exhibit 3.10(b) displays the effects of a subsidy. In both diagrams, the downward sloping line labeled *D* and the upward sloping line labeled *S* are the same demand and supply curves that we have been working with in Exhibits 3.8 and 3.9. Given *D* and *S*, the free-market equilibrium values for price and quantity continue to be $P^*$ and $X^*$.

Sales taxes and subsidies are typically specified as percentage amounts (we are all familiar with, for instance, a 3 percent sales tax). In Exhibit 3.10, we for simplicity and with no loss of generality represent both the tax and the subsidy as a flat dollar amount per unit of *X*. This enables us to represent the impact of the tax and of the subsidy in our analysis as simple parallel shifts of the demand curve with the dollar amount of the tax or subsidy determining the size of the vertical shift. The tax shifts the demand curve down to the line labeled $D^{TAX}$ [Exhibit 3.10(a)], and the subsidy shifts the demand curve up to the line labeled $D^{SUB}$ [Exhibit 3.10(b)].

In Exhibit 3.10(a), the intersection of the demand curve $D^{TAX}$ and the supply curve $S$ establishes the equilibrium quantity, $X^{TAX} < X^*$. At this new (and reduced) quantity, consumers pay the price labeled $P^{PAID(TAX)}$ and sellers receive the lower price labeled $P^{RECEIVED(TAX)}$. Note that

- $P^{PAID(TAX)} > P^* > P^{RECEIVED(TAX)}$.
- $P^{PAID(TAX)} - P^{RECEIVED(TAX)}$ is the flat dollar tax per unit sold.

In Exhibit 3.10(b), the intersection of the demand curve $D^{SUB}$ and the supply curve $S$ establishes the equilibrium quantity, $X^{SUB} > X^*$. At this new (and larger) quantity, consumers pay the lower price labeled $P^{PAID(SUB)}$, and sellers receive the higher price labeled $P^{RECEIVED(SUB)}$. Note that

- $P^{PAID(SUB)} < P^* < P^{RECEIVED(SUB)}$.
- $P^{RECEIVED(SUB)} - P^{PAID(SUB)}$ is the flat dollar subsidy per unit sold.

Here are the comparative statics. With the tax, the quantity exchanged in the market is less, consumers pay more, sellers receive less, and government has a revenue source in the amount of $X^{TAX}(P^{PAID(TAX)} - P^{RECEIVED(TAX)})$. With the subsidy, the quantity exchanged in the market is more, consumers pay less, sellers receive more, and the program is a tax drain for the government in the amount of $X^{SUB}(P^{RECEIVED(SUB)} - P^{PAID(SUB)})$. Is the sales tax a good way for the government to raise revenues and/or to control output? Is the subsidy a good way for the government to encourage more output and consumption of the good $X$?

It is worthwhile to think further about how all of this plays out, about what all of the consequences are and, ultimately, about how desirable the legislation is. The imposition of a sales tax for the sole purpose of raising revenue may certainly be justified by a government's need for funds (of course a sales tax may or may not be the best form of taxation but, regardless, tax revenues are needed). Aside from this, microeconomists typically do not favor these interventions in the workings of the free market, and our discussion certainly suggests why they do not. But these issues have been debated for decades, and intervention goes back a good deal further than that. Is it because there are other, noneconomic considerations? Is it because economic theory that is based on the assumptions that markets are frictionless and that competition is vibrant overlooks certain real-world considerations, which, if taken into account, could lead to different conclusions? It is tempting at this point to come up with a few normative "shoulds." But, for now, we will leave the matter for further thought and discussion, and will progress with our analysis.

## EQUILIBRIUM IN THE MARKET FOR SHARES OF THE RISKY ASSET

We next apply demand-supply analysis to the market for the risky asset. Our analysis builds on the discussion presented earlier in this chapter where we obtained the investor's downward sloping demand to hold shares of the risky asset. The individual demand curves can be horizontally aggregated to obtain the market's demand curve, just as we did for the generic good, $X$.

Given the market demand curve, how is price set on the market? What is missing? A supply curve, of course . . . or maybe not.

In this analysis, the supply of shares is a fixed number, it is fixed by the number of shares of the risky asset that have been issued and are outstanding. In Exhibit 3.11(a) we show a market demand curve labeled $D^{MKT}$ and a vertical (zero elasticity) curve labeled $S$ that intercepts the horizontal axis at the value $N^*$ (the number of shares outstanding). The demand curve intersects the vertical line at the price labeled $P^*$. This price is the equilibrium value of shares. At a higher price, the market in aggregate would wish to hold fewer than $N^*$ shares because the expected return and its attending risk premium would be lower. At a lower price, the market in aggregate would

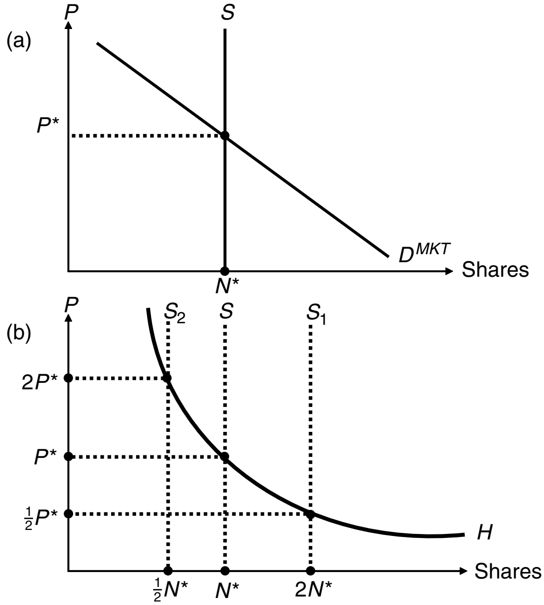


**EXHIBIT 3.11** The Market Demand Curve to Hold Shares of a Risky Asset [3.11(a)], and a Rectangular Hyperbola ($H$) That Shows How Price Changes as the Number of Shares Outstanding Change [3.11(b)]

wish to hold more than $N^*$ shares because the expected return and its attending risk premium would be higher. But there are only $N^*$ shares around to be held and so, consequently, the equilibrium price is $P^*$.

Even though the vertical line in the diagram is labeled "*S*," we have been careful not to refer to it as a supply curve, for a reason that you will soon understand. Typically the demand and supply curves for the generic good, *X*, are independent of each other. If demand shifts, a new equilibrium is found on the supply curve. If supply shifts, a new equilibrium is found on the demand curve. Recognizing this, we can define the supply curve as the set of equilibrium points that are traced out when demand shifts, supply constant. Similarly, we can define the demand curve as the set of equilibrium points that are traced out when supply shifts, demand constant.

This will not work in the current setting. Let's figure out what curve would be traced out as the vertical line labeled *S* shifts. If the *S* curve shifts (if $N^*$ changes), earnings per share must also change. For instance, if a firm has a 2 for 1 stock split, the number of shares outstanding will double and earnings per share will be cut in half. What effect should this have on share price? All else equal, price will also be cut in half. In Exhibit 3.11(b) we show the curve that would be traced out by changes in the number of shares outstanding. That curve is labeled *H*. The *H* curve passes through the $P^*$, $N^*$ point identified in Exhibit 3.11(a). If a 2 for 1 stock split occurs, the *S* curve with an intercept on the quantity axis of $N^*$ shifts to $S_1$ with a quantity axis intercept of $2N^*$, and price falls to $(1/2)P^*$. Or if a 1 for 2 reverse split occurs, the *S* curve shifts to $S_2$ with a quantity axis intercept of $\frac{1}{2}N^*$, and price rises to $2P^*$. But note, the total value of shares does not change as these shifts occur. That is, all else equal,

$$P^* N^* {}^1\!/_4 {}^1\!/_2 \, P^* 2N^* \; {}^1\!/_4 \, 2P_1^* = {}_2N^*$$

In other words, as we move along the *H* curve, price times the number of shares outstanding remains constant. This curve, quite clearly, is not a demand curve in any meaningful sense of the term.

The *H* curve that we have just described is a *rectangular hyperbola*. Because the total value of share holdings is constant along the rectangular hyperbola, the price elasticity of demand at every point along the curve is unity. A demand curve that is preordained to be a rectangular hyperbola does not make for a very interesting demand curve, however. Having identified it, let's move on. We will continue our analysis with reference to the demand curve $D^{MKT}$ shown in the top diagram. But as we proceed, we will not let the *S* curve shift.

The next step in our analysis is to show how trading takes place in the frictionless environment when the market demand curve, *D*, shifts. To do this, we first return to the demand curve of an individual investor. This curve is displayed in Exhibit 3.12(a).

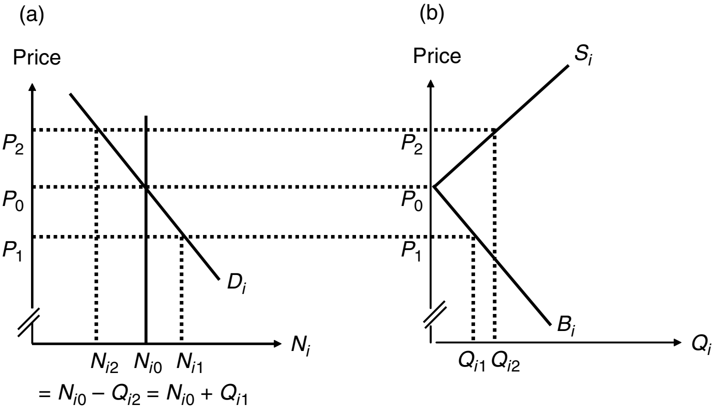


**EXHIBIT 3.12** The $i^{th}$ Investor's Demand Curve to Hold Shares [3.12(a)] and the Associated Buy and Sell Curves for That Investor [3.12(b)]

In Exhibit 3.12(a) the symbol "$D_i$" identifies the downward sloping demand curve of our $i^{th}$ representative investor. In that diagram, the $i^{th}$ investor's current holdings are identified by the vertical line at $N_{i0}$. Do not confuse the vertical line at $N_{i0}$ with the vertical line at $N^*$ in Exhibit 3.11. $N_{i0}$ is the (variable) number of shares held by the $i^{th}$ investor; $N^*$ is the total (fixed) number of shares outstanding for the asset. We denote the $i^{th}$ investor's alternative share holdings that we have shown by subscripts 0, 1, and 2. Similar subscripts are used for the price variable. $P_0$ is the price at which this investor is willing to hold $N_{i0}$ shares of the market portfolio. At any price greater than $P_0$, the participant would like to hold fewer shares; at any price lower than $P_0$, he or she would like to hold more shares.

If price in the frictionless market is $P_1$, the investor will want to hold $N_{i1}$ shares, and he or she will seek to achieve this by buying $Q_{i1}$ shares. If, instead, the price is $P_2$, the investor will want to hold $N_{i2}$ shares, and he or she will seek to achieve this by selling $Q_{i2}$ shares, and so forth. From the comparative statics shown in Exhibit 3.12(a), we can obtain the investor's *buy and sell curves*, which are labeled $B_i$ and $S_i$, respectively, as shown in Exhibit 3.12(b). The buy curve, $B_i$, plays the role of a demand curve, and the sell curve, $S_i$, plays the role of a supply curve. But in point of fact, the investor truly has only one demand curve, the demand curve to hold shares. [$D_i$, as shown in 3.12(a)].

Note the following about the demand curve to hold shares and the associated buy/sell order curves:

- The buy/sell functions for the $i^{th}$ investor branch off of the price axis at a value determined by the intersection of that investor's demand curve

$(D_i)$ and the vertical line at $N_{i0}$ that denotes the number of shares that the investor is currently holding.

- Holding $N_{i0}$ constant, the point of intersection with the vertical line at $N_{i0}$ depends on the location of the demand curve. Accordingly, shifts in $D_i$ are associated with shifts in the buy/sell order curves.
- Holding $D_i$ constant, the point of intersection with the vertical line at $N_{i0}$ depends on the current size of the $i^{th}$ investor's share holdings. Therefore, each trade is accompanied by *a shift* of the buy/sell order curves (the curves shift down with a purchase and up with a sale). For instance, if the investor were to purchase $Q_{i1}$ shares, the price intercept for $B_i$ and $S_i$ would shift down from $P_0$ to $P_1$. Or, if the investor were to sell $Q_{i2}$ shares, the price intercept for $B_i$ and $S_i$ would shift up from $P_0$ to $P_2$.

## MARKET EQUILIBRIUM

When the market for the risky asset is out of equilibrium, the market regains equilibrium as investors trade with each other so as to adjust their individual share holdings to a market price that just balances their aggregate buy and sell desires. To see how this works, assume an initial distribution of shares among the participants (the $i^{th}$ investor holds $N_{i0}$ shares) that reflects the results of a prior round of trading. In the frictionless market that we have thus far been assuming, the prior round of trading could have occurred in the previous, infinitesimally close instant in time. Regardless, assume that something suddenly happens that shifts the individual demand curves so that a new equilibrium price has to be established.

That "something" could be the advent of new information, or a shift in investors' tastes for risk, or a change in their wealth positions and/or need for cash, or some other factor that caused a change of their expectations. Following this change, a new price is set based on the orders that the investors submit to the market. Because the market is frictionless, all investors submit their orders instantaneously and simultaneously. Furthermore, their orders are not simple "buy 100 shares at \$32"–type orders—they are each participant's complete buy and sell order functions. These complete order functions for the full set of investors are instantaneously aggregated and matched, a new equilibrium price is established, the shares are redistributed across the investors according to the trades that have been consummated, and everybody's desire to adjust their holdings of the risky asset is simultaneously and instantaneously satisfied. What a wonderful world! Let's see how it works.

We have just noted that, given each investor's order functions (the $S_i$ and $B_i$ trade curves) and an environment where trading is costless, each investor submits his or her complete trade function to the market for one round of trading. The reason is threefold. First, the order size associated with each price along the trade curves will have been accurately written given the price that is established on the market. That is, each point along the investor's demand curve can be viewed as an if statement: For instance, "if the price turns out to be $P_1$, than I wish to buy $Q_{i1}$ shares; or, if the price turns out to be $P_2$, than I wish to sell $Q_{i2}$ shares; and so on." Consequently, if the price actually turns out to be $P_1$, the investor will in fact buy $Q_{i1}$ shares, and the mission will have been accomplished to the investor's satisfaction.

The second reason why each investor will submit his or her complete trade curves to the market is that the instantaneous market clearing procedure that we have assumed for the frictionless trading environment guarantees that the investor realizes only one execution (i.e., the investor will get an execution at only one price, the market clearing price). The third reason is that no individual trader is large enough relative to the market to have the power to affect the price that is established by the market.

Aggregation from individual buy and sell order curves to the aggregate market curves follows the standard procedure that we have already established: namely, aggregate all of the quantities (i.e., sum up the individual buy orders horizontally at each price, and do the same with the individual sell orders). The aggregated, market buy and sell curves are shown in Exhibit 3.13. They are labeled $B^{MKT}$ and $S^{MKT}$, respectively.

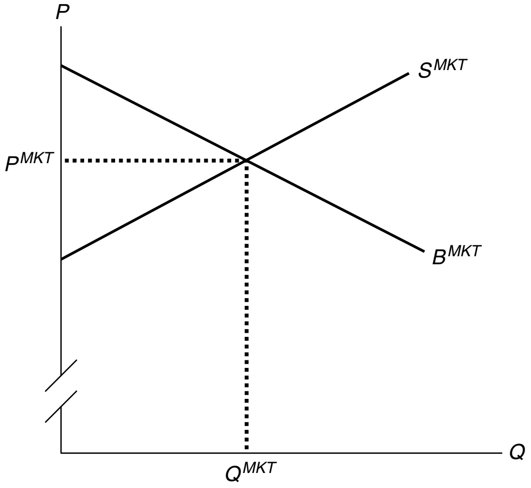


**EXHIBIT 3.13** Determination of Market Values for Price and Number of Shares Traded ($Q^{MKT}$) by the Intersection of the Market Buy Curve ($B^{MKT}$) and Sell Curve ($S^{MKT}$)

Unlike the $B_i$ and $S_i$ curves for an investor that are shown in Exhibit 3.12, the market curves $B^{MKT}$ and $S^{MKT}$ can intersect each other because the locations of the $D_i$ curves and the vertical lines at $N_{i0}$ vary across investors. When the aggregate buy and sell trade curves intersect, prices exist at which some investors are willing to sell, and other investors are willing to buy. Hence, trades can occur. For the configuration shown in Exhibit 3.13, $Q^{MKT}$ shares will change hands at a price of $P^{MKT}$.

Exhibit 3.14 presents further detail on how the equilibrium shown in Exhibit 313 is achieved. The exhibit does this by presenting four diagrams that show the effect that the exchange of the $Q^{MKT}$ shares at a price of $P^{MKT}$ has on the $i^{th}$ investor's demand curve to hold shares. Trading eliminates orders from the market's buy and sell curves. After the subtraction of $Q^{MKT}$ shares (they have traded, they are gone) from both the $B^{MKT}$ and $S^{MKT}$ curves, the market's trade curves will have shifted to the left until the intersection point shown in Exhibit 313 is at the price axis, as shown in Exhibit 3.14(a).

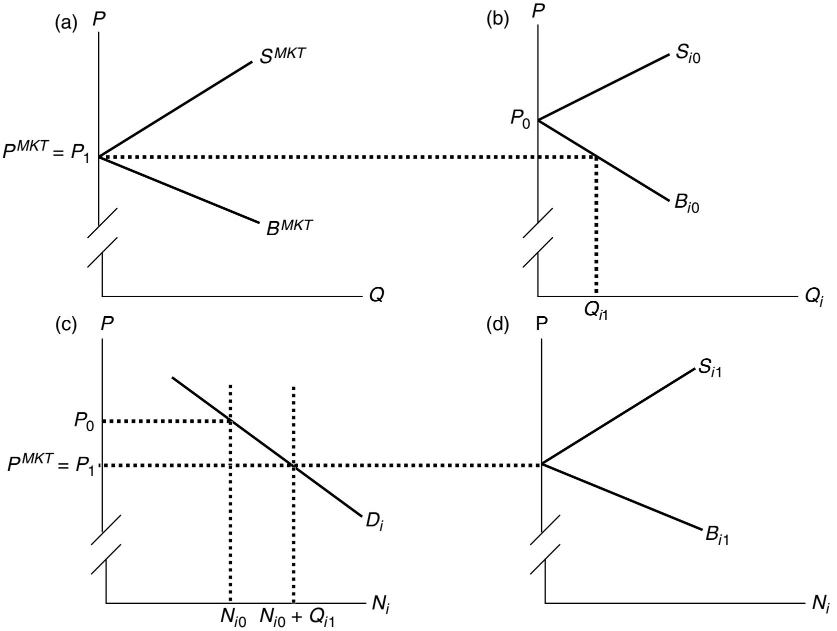


**EXHIBIT 3.14** Market Buy and Sell Curves [$B^{MKT}$ and $S^{MKT}$] after Trades Have Been Made at $P^{MKT} = P_1$ [3.14(a)], Determination of Individual investor's Transaction (Bought $Q_1$ shares at $P_1$) [3.14(b)], Individual Investor's Demand Curve after Having Bought $Q_1$ Shares at $P_1$ [3.14(c)], and Individual Investor's Buy and Sell Curves after Having Bought $Q_1$ Shares at $P_1$ [3.14(*d*)]

The process of trading harmonizes the $i^{th}$ decision maker's investment desires with the market. As shown in Exhibit 3.14(b), the trade curves the $i^{th}$ investor has submitted to the market results in the purchase of $Q_{i1}$ shares at the price $P^{MKT}$. Accordingly, Exhibit 13.4(c) now shows that, after the trade, the vertical line that denotes the $i^{th}$ investor's holdings has shifted from $N_{i0}$ to $N_{i0} + Q_{i1}$. Consequently, the vertical line now intersects $D_i$ at the market price, $P^{MKT}$. Thus, after the trade, the $i^{th}$ investor is holding exactly the number of shares that he or she would like to hold at the current market price.

If the investor were to submit a new set of trade curves to the market after the trade, the curves would appear as shown in Exhibit 3.14(d). Notice that the new trade curves for the $i^{th}$ investor necessarily branch off the price axis at $P^{MKT}$. Because this is true for each and every trader, mutually profitable trading opportunities no longer exist immediately after a trading session for any pairing of market participants (nobody is looking to buy shares at a price higher than anyone else is looking to sell them). Therefore, one round of trading in the frictionless environment harmonizes the trading propensities of all investors, and there will be no recontracting (i.e., there is no additional trading).

The market's aggregate demand curve to hold shares of the risky asset can now be identified: It is the curve labeled $D^{MKT}$ in Exhibit 3.15(b). The relationship between the market's trade curves, $B^{MKT}$ and $S^{MKT}$ [as shown in Exhibit 3.15(a)], and the market's demand curve, $D^{MKT}$, is the same as that which relates $B_i$ and $S_i$ to $D_i$ (as shown in Exhibit 3.12).

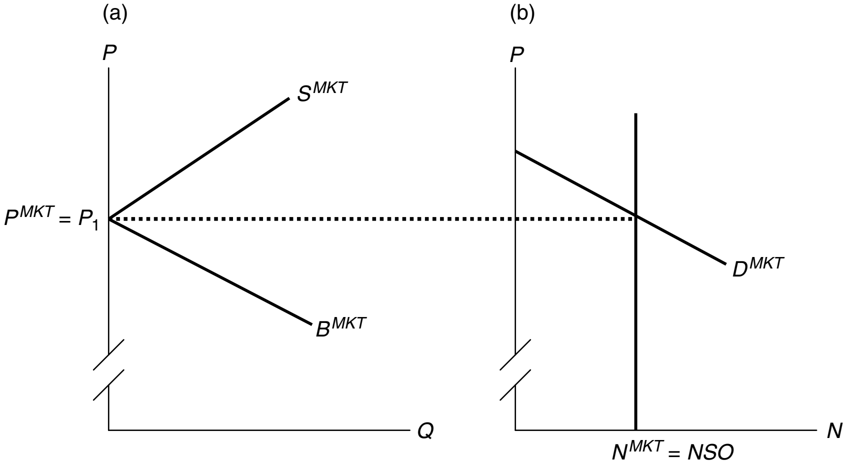


**EXHIBIT 3.15** Market Buy and Market Sell Curves [3.15(a)] and the Market Demand Curve to Hold Shares [3.15(b)] after Trades Have Been Made at $P^{MKT}$

Unlike the case for an individual, the market in aggregate must hold a *given* number of shares—the aggregate *number of shares outstanding* (NSO). Thus we take the vertical line at NSO as fixed. In equilibrium, $D^{MKT}$ crosses the vertical line at $P^{MKT}$, the price that has been established on the market. Alternatively stated, the equilibrium price for the market can be obtained by assessing the market demand equation at $D^{MKT} = NSO$. Because the market in aggregate must hold the total number of shares outstanding, and because this number is given, a separate supply equation is not needed to obtain a solution. There are two equations ($D^{MKT} = f(P)$ and $D^{MKT} = NSO$), two unknowns ($D$ and $P$), and we can solve for their equilibrium values.

As we have already said and will now repeat, the vertical line at $NSO$ is not a supply curve in the following sense. If total corporate earnings, dividends, growth, and so on, are unaffected by the number of shares outstanding, then any change in $NSO$ (due, for example, to a stock split, stock dividend, or reverse split) would be associated with an equal but opposite percentage change in share price. This being the case, with demand propensities constant, shifts of the vertical line at $N^*$ would trace out a locus of equilibrium prices, which lie on the negatively inclined, convex curve of unitary elasticity that we previously noted is a rectangular hyperbola. It would be misleading to consider this curve the market demand curve for the risky asset. Therefore, NSO should not be considered a supply curve.

## STOCKS, CAKES, AND SUBSTITUTABILITY

Demand and supply analysis is important because price is an important rationing variable. That is, it is the impersonal working of prices in the marketplace that, in keeping with Adam Smith's invisible hand, directs the allocation of inputs (into production) and outputs (for consumption) among their competing uses. It is important to repeat this: in a competitive, free-market environment, it is price, not the edicts of some public administrator, that directs the flow of resources.

Price is an important rationing variable, however, only if the flow of resources responds to price changes. For this reason, elasticity is a concept of number-one importance. We have treated it as such in this chapter, and we will end this chapter with one more discussion of it. In so doing, we will return to the demand curve to hold shares of the risky asset.

We have referred to the risky asset as the "risky market portfolio." The demand to hold shares of the risky market portfolio is responsive to price because there is a substitute asset to hold, the risk-free asset. The less risk-averse an investor is, the closer the market portfolio and the risk-free asset are as substitutes for that participant. We have seen that, in the limit, if the investor is totally risk-neutral, the two assets are perfect substitutes and the investor's demand for either of them is infinitely elastic at the current price for the market portfolio that, given $E(P_T)$, equates the expected return on the risky market portfolio with $r_f$.

We discussed this with reference to Exhibit 3.6, and repeat the thought here to give ourselves a platform for raising the final question that we address in this chapter: if the demand to hold shares of the market portfolio is price-elastic (but not infinitely elastic) for a risk-averse investor, how elastic is the investor's (and the market's) demand for any of the individual stocks that comprise the market portfolio?

An answer to this question requires knowledge of how individual stocks are priced as they are combined to form the risky market portfolio. This topic is commonly addressed in finance courses, and the Capital Asset Pricing Model provides an excellent answer for a frictionless marketplace that is characterized by homogeneous expectations. In brief, this is how it works.

For simplicity, let all investors have the same investment horizon (e.g., one year). The one-year returns for each individual stock are assumed to be normally distributed, and the normal distribution can be described by a mean and a variance. Standard deviation, the square root of variance, is our commonly accepted measure of risk. A stock's risk can be decomposed into two components. One component is driven by factors that affect all stocks in the broad market; accordingly, it is called *systematic risk*. The other component is unique to an individual stock; accordingly, it is called *unsystematic risk*. Different stocks can have different exposures to the common market factors and thus different levels of systematic risk. For example, a change in interest rates will have broad effects throughout the economy, but the impact will be greater for firms such as commercial banks whose business is particularly interest rate sensitive. So, here is where we stand:

**Systematic risk:** To some degree, a stock's return is correlated with the return on the market portfolio. Common information factors such as change in a macroeconomic variable (interest rates or aggregate GDP to name two) can account for this correlation. To the extent that a stock's returns are correlated with the market, its inclusion in the market portfolio contributes to the riskiness of the market portfolio. Alternatively stated, the riskiness of the market portfolio is attributable to the inter-correlated risk components of all of the individual stocks that together comprise the market portfolio. This is why the first component is called *systematic risk*. It is also called *nondiversifiable risk*.

**Unsystematic risk:** To some degree, a stock's return is uncorrelated with the market. That is, some portion of its price fluctuations is attributed to factors that are unique to the firm (for instance, the approval of a bank loan, the advent of merger news, or a scientific breakthrough that the firm has just made). The component of risk that is uncorrelated with the market (the stock specific variability) does not contribute to the risk of the market portfolio, which is why it is called nonsystematic risk. It is also called *diversifiable risk*.

The second component of risk is referred to as *diversifiabl*e because price changes that are independent across firms tend to cancel each other out. That is, one firm can be experiencing a positive shock (good news) while some other firm is experiencing a negative shock (bad news). This is what portfolio diversification is all about: idiosyncratic (nonsystematic) factors can be rendered inconsequential by sufficient portfolio diversification (that is, as we have said, independent price changes tend to cancel each other out). In contrast, the first component is not diversifiable because stocks that are positively correlated with the risky market portfolio move in harmony with each other to the extent that they are affected by the same factors (and thus for this component of risk there are no independent moves to cancel each other out).

And so, what matters to a sufficiently diversified investor is not the total risk (variability) of a stock's return, but only systematic risk, the first of the two components. Accordingly, each individual stock is priced so that its expected return offers a risk premium that compensates investors for the systematic risk which the individual stock contributes to the market portfolio.

We can be more specific with regard to nondiversifiable risk. This risk, for a stock, is measured by the stock's *covariance with the market*. A covariance is essentially a kind of variance. Think of it this way. The variance for stock $i$ can be written as

$$variance_i = \sigma_i^2 = \sigma_i \sigma_i \rho_{i,i} \tag{3.21}$$

where $\sigma_i$ is the $i^{th}$ stock's standard deviation and $\rho_{i},_{i}$ is the $i^{th}$ stock's correlation with itself. Clearly, a stock is perfectly correlated with itself and thus $\rho_{i,i} = 1$. While a stock's variance in essence tells us how its return *varies with itself*, a stock's covariance with another stock or portfolio tells us how the stock's return varies with the returns on the other asset. A slight modification of equation 3.21 gives us the $i^{th}$ stock's covariance with the risky market portfolio:

$$covariance_{i,m} = \sigma_i \sigma_m \rho_{i,m} \tag{3.22}$$

The covariance is not a variance because, in general, $\sigma_i \neq \sigma_m$ and $\rho_{i,m} \neq 1$.

Equation 3.22 is what we need to carry on. Intuition and the Capital Asset Pricing Model tell us that a stock is priced according to its covariance with the market ($cov_{i,m}$). Alternatively stated, in the frictionless environment, the risk premium that must be paid on a stock (and thus its current price in the marketplace) depends on just two things: the risk-free rate and its covariance with the market.

So, what does this have to do with cakes? Let's circle back to our two-cake example and find out. Those two cakes (A and B) are baked according to recipes that say how the ingredients for each should be combined, and the two cakes are substitutes for each other. A risky market portfolio can similarly be formed; the individual stock covariances are the ingredients, and the percentage weights for each of the stocks is the recipe. In other words, the optimal portfolio is identified by the percentage weights given to each of the individual stocks that comprise it, much as a cake is identified by its ingredients and how they are combined (i.e., its recipe).

Let's first consider how the substitutability between cakes compares with the substitutability between portfolios. The choice between cakes A and B depends on each individual's own preferences and, while everyone may consider the two cakes as substitutes for each other, people facing identical prices for the two cakes will make different selections according to their individual tastes. Consequently, neither cake will totally dominate the other in consumers' shopping baskets (or ovens). Two different risky portfolios with different risk and return configurations may likewise be considered substitutes for each other and, as with the two cakes, one might presume that some investors will select one risky portfolio while others will select the other risky portfolio.

Not so when a risk-free asset is introduced. Take another look at Exhibit 2.8. Without the risk-free asset, some investors may select portfolio 1 (those with the least risk aversion), others may select portfolio 3 (moderate risk aversion), and yet others may select portfolio 2 (high risk aversion). But with the risk-free asset and the ability to form a portfolio combination of the risky portfolio and the risk-free asset, portfolio 3 will dominate portfolios 1 and 2 for everybody, regardless of any single investor's individual taste for risk. So, in a CAPM world, there is but one risky market portfolio. Namely, all investors will choose the dominating risky portfolio regardless of their tastes for risk (as long as they are risk-averse, which we assume them to be). Consequently, in a CAPM world there are just two cakes: the risky market portfolio and the risk-free asset.

Let's now consider the composition of cakes and portfolios. The ingredients of a given cake are perfect complements in that the recipe for making it calls for specific proportions (that is, you cannot change the proportions without changing the cake). Each and every single stock in the risky market portfolio, on the other hand, not only has substitutes—it has perfect substitutes. Here's why.

It can be shown that the covariance for any stock (call it stock A) can be replicated exactly by properly combining two other stocks (call them stocks B and C). This means that a proper combination of B and C is a *perfect substitute* for A. Do you recall what an individual's demand curve for *X* looks like when *Y* is a perfect substitute for *X*? Under the condition of perfect substitutability, *X*'s demand is infinitely elastic at its appropriate price. Likewise for stock A: Its demand is infinitely elastic at its appropriate price, given its covariance with the market. Stocks with higher covariances with the market add more to aggregate market risk; accordingly, they must offer higher risk premiums and, to do so, they have to trade at lower prices. Stocks with lower covariances add less to aggregate market risk; accordingly, they must offer lower risk premiums and, to do so, they have to trade at higher prices.

At this point, combining stocks in a portfolio, and combining butter, eggs, flour, and so on in a cake, may appear to be polar opposites. The ingredients required to bake a specific cake are perfect complements, and they have to be combined according to the precise recipe for the particular cake. The stocks required to form the risky market portfolio are perfect substitutes for each other and, if any one stock is not properly priced it will not be held at all (if its price is too high) or will be oversubscribed (if its price is too low). So, while the demands for the ingredients for a cake are totally price-inelastic (as long as that cake is the one being baked), the demand to hold shares of any individual stock is infinitely elastic at one price—the price at which that stock must trade. That is why CAPM is the Capital Asset *Pricing* Model.

The fundamental difference between the cakes and the portfolios is that the ingredients of the risky market portfolio are perfect substitutes for each other while the ingredients for the cakes are not. Sugar is different than flour, which is different than butter. Stock A, we have seen, is not different than a replicating combination of stocks B and C. It is as if we were dealing with two different tubs of butter, one produced in Vermont and the other produced in Wisconsin: The butter is the same and it is butter that is being priced. So what is it that makes two different stocks (or portfolio combinations of stocks) the same? They are the same if their covariances with the risky market portfolio are the same. Again, we emphasize that only covariance matters.

So, in a CAPM world, a stock's price is based on just one single variable, its covariance with the risky market portfolio. This is how the forces of supply and demand work: As we have already noted, any stock that trades at too high a price given its covariance will not be held (nobody will want to buy it) and thus its price must fall, while any stock that trades at too low a price given its covariance will be over-subscribed and thus its price must rise. Given the risk-free rate, the set of stocks, and their covariances, we know the price that each stock must trade at and the composition of the risky market portfolio.

An important theorem in financial economics captures this thought. It is called the *separation theorem*. The separation theorem states that the optimal combination of different stocks in the risky market portfolio is the same for all investors, regardless of their individual risk, return preferences (just as there is only one way to bake any given cake). As we have said, in a CAPM world, there are just two "cakes": the risky market portfolio and the risk-free asset. In the case of cakes, individual preferences enter the picture only when deciding how much of cake A to consume relative to cake B; in the case of risky assets, individual preferences enter the picture only when deciding how much to invest in the risky market portfolio, and how much to invest in the risk-free asset. Separation of the specification of the risky market portfolio from individual investment decisions (i.e., how much to invest in the risky market portfolio and how much to invest in the risk-free asset) explains why the separation theorem has been given this name.

We will say it one more time: In a frictionless world characterized by homogeneous expectations, it is not the individual stocks per se that are priced, it is risk that is priced. The power of CAPM is that it shows rigorously how the price of risk translates into the price of individual stocks in accordance with just this one thing: their systematic risk. Is this useful to know? The answer is a solid, resounding "yes." Portfolio management should and has taken account of this CAPM result. But how well does the CAPM model fit with real-world, risky asset markets in other respects? How "collectively" realistic is it?

With this question concerning the realism of a model in mind, we suggest that you flip back to the second question in Chapter 1's "Questions to Think About," where the issue of realistic assumptions and realistic theories is raised. CAPM is based on a specific set of assumptions, and we have stated the key ones. They include:

- All investors are risk-averse.
- All investors have the same investment horizon (e.g., one year).
- Returns are normally distributed (thus mean and variance are the only parameters to consider).

- All investors can borrow and lend unlimited amounts at a constant risk-free rate ($r_f$).
- All investors are in possession of the same information.
- Investors all have the same (homogeneous) expectations based on that information (they all agree about the mean and variance parameters of the returns distribution).
- Investing/trading is a costless (frictionless) process.

In Question 2, we ask, "How do you feel about these assumptions? Could a set of assumptions that are individually unrealistic, collectively have descriptive/predictive power? What is the alternative if you do not make some of these assumptions?"

How would you answer these questions now? As you give this more thought, keep the following in mind: the assumptions that one makes depends on the questions and issues that one wishes to address. For our purposes, it is now time to leave the frictionless world and to consider more closely the operations of real-world markets. We head in that direction in Chapter 4.

## CHAPTER SUMMARY

Our number-one objective in this chapter has been to understand how price and quantity are set in a marketplace that comprises a large number of buyers and sellers (large enough, that is, so that no one participant has any significant presence in the market and no one, therefore can affect a market determined price). In addressing this issue, we have continued to assume that the marketplace is a frictionless environment and, in so doing, we have built on the consumer choice model presented in the previous chapter. The following steps that we have taken are key:

- We extended the comparative static results of Chapter 2 to obtain the decision maker's demand curve for a good or service, and for a financial asset. In so doing, we specified the relevant price for the economic resource under consideration, and the set of other variables that are constant along a demand curve. Throughout, we have taken tastes (the decision marker's utility function) as constant and, for simplicity, have assumed a two-good universe. For a generic product, the demand curve shows how $X$ (the quantity of the product demanded) depends on $P_X$ (the price of $X$), the price of the other generic product ($P_Y$), and the decision maker's money income ($M$). A change in $P_X$ (all else constant) causes a movement along the demand curve, while a change in either of the other two demand-determining variables causes the demand curve for $X$ to shift.

- Elasticity was introduced as the measure of $X$'s responsiveness to each of the independent variable (with the other two variables being held constant). We identified three elasticities: own price elasticity (the responsiveness of $X$ to its own price), cross-price elasticity (the responsiveness of $X$ to the price of the other product), and income elasticity (the responsiveness of price to the income variable). Complementarity and substitutability between $X$ and $Y$ were discussed with reference to $X$'s own price elasticity, and to $X$'s cross-price elasticity with respect to $Y$. Clearly, relative prices are of greater importance to the decision maker, the better the substitutability between the two goods.
- The demand curves of individuals were aggregated to obtain the demand curve for a market. Aggregation is a simple process: sum up the quantities demanded by each of the individual participants at each of the different values of price.
- The demand analysis was next applied in a more defined context. Building on the risk and return formulation presented in Chapter 2, we obtained the demand curve to hold shares of a risky asset in a two-asset environment (the risky asset and a risk-free asset). The risky asset was viewed as a market portfolio that comprises the full set of individual stocks. Unlike the generic good, $X$, which is flow-dimensioned, the demand to hold shares of the risky market portfolio is stock-dimensioned. We worked in the context of a single period (e.g., one year) model. The risk-free rate of interest $[r_f]$ and the expected end of period price of the market portfolio $[\mathrm{E}(P_T)]$ constant, the demand to hold shares of the risky asset is a negative function of the current share price ($P$) of the risky asset. This is because, as $P$ decreases, the risk premium, $\frac{E(P_T)}{P} - 1 - r_f$ rises and thus a risk-averse investor will seek to hold a larger number of shares.
- Returning to the generic product, $X$, we matched the market demand curve for $X$ with a market supply curve of $X$ to obtain the equilibrium values of price and quantity. The important properties of a supply curve were established and the intersection of a downward sloping demand curve with an upward sloping supply curve was shown to establish a stable equilibrium price and quantity of $X$.
- We then saw how regulatory intervention in the market for $X$ can distort the free-market price and quantity equilibrium. In particular, we considered the effects that can attend a government imposed price ceiling, or a price floor, or a sales tax or a subsidy.
- Returning to the market for shares of the risky asset, we applied demand-supply analysis to obtain equilibrium in the marketplace where

shares of the risky asset are traded. We clarified that in this setting there is only one demand curve: the demand to hold shares of the risky asset. From this curve, for an individual, we obtained two other curves: a downward sloping curve to buy shares at lower prices, and an upward sloping curve to sell shares at higher prices. It was the individual buy and sell curves that we summed across investors to get aggregate market buy and sell curves, and we matched these two aggregates to establish the number of shares that would trade, and the price that they would trade at in the market.

- We concluded the chapter with the thought that, in the frictionless environment, it is not individual shares that are priced, it is risk that is priced. But one does not buy risk. Rather, the market pays a participant for accepting it. In the perfect, frictionless environment, perfect substitutes exist for every individual stock, and the market's demand curve to hold shares of any individual stock is infinitely elastic at a price that depends on the stock's covariance with the market. This is because only systematic (as distinct from diversifiable) risk is priced (diversifiable risk can be reduced to nil by sufficient portfolio diversification), and a stock's covariance with the market measures the stock's systematic risk, its contribution to the riskiness of the market portfolio (which is our risky asset).

## LET'S CONSIDER AN APPLICATION: BANDWAGON AND SNOB EFFECTS

The section ‘‘Let's Consider an Application’’ concerning philanthropy at the end of Chapter 2 focused on a utility maximizing consumer's decision to give away part of his or her income. The economic rationality behind this act was explained by introducing the concept of an interrelated utility function. Namely, the decision maker's utility is a function not only of the goods and services personally consumed, but also of the utility achieved by others. In this thought question, we focus on another way in which the decision maker can be affected by other people. Rather than caring about others per se, the utility a decision maker gets from consuming a particular good or service depends on the amount of that good or service that is consumed by others. In other words, the decision maker's tastes (read ‘‘utility function’’) depend on what others do. Let's think about how this might work.

Consider the good called "*X*." What are your tastes for *X*? Neat little item, you might think, but the price is a bit high. Maybe I will not buy much of it right now. But then you see that more people are buying *X*. Hmm, *X* is in vogue. Perhaps I will go ahead with that purchase even though the price is high, our decision maker thinks. We call this positive dependence on other people's decisions a *bandwagon effect*. It is akin to a *momentum effect*, or to *herding*. It is one form of interpersonal dependency.

Here is another form of interpersonal dependency. Consider your feelings about a good called "*Y*." Hmm, *Y* is another neat little item to get, you think, and the price seems fine. And then, before you make the purchase, you see that other people are flocking to the product and that turns you off. *Y* is seen all over the place; I do not want to follow the crowd and be like everyone else, the decision maker thinks. For now, I will hold back on my purchase of *Y*. We call this negative dependence on other people's decisions a *snob effect*. It is akin to *reversal behavior*.

With this by way of background, we invite you to do the following.

1. Think of a few examples of consumer goods for which a bandwagon or a snob effect might be observed.
2. Assess the consumption of *X* in terms of the number of units of *X* consumed per unit of time (that is, treat the quantity of *X* demanded as a flow variable), and assume that a representative individual has a downward sloping demand curve to consume *X*. Do the same for *Y*. Think about why we want to make this assumption.
3. Consider how momentum and reversal effects might characterize trading in the equity markets.* In what way is a momentum move akin to a bandwagon effect, and reversal behavior akin to a snob effect?
4. Answer the following question on the basis of your intuition: Does the bandwagon effect make the market demand curve for *X* more or less price-elastic?

*In an equity market a momentum move is a series of price changes in the same direction (i.e., up-ticks followed by more up-ticks, and down-ticks followed by more down-ticks). A reversal move, on the other hand, is down-ticks followed by up-ticks and up-ticks followed by down-ticks.

(*Continued*)

(*Continued*)

5. Answer the following question on the basis of your intuition: Does the snob effect make the market demand curve for $Y$ more or less price-elastic?
6. Graphical analysis can be used to obtain a more rigorous answer to questions 4 and 5. We will describe the steps to take and ask you to implement them. First do the analysis for a bandwagon effect.
   - **a.** Draw a downward sloping demand curve for an individual for the product $X$. Since the individual's demand for $X$ depends on the amount of X that others consume, label this demand curve $X^{MKT} = X_1$, where $X_1$ is one specific value of the total market quantity of $X$ being consumed.
   - **b.** Aggregate the individual $X^{MKT} = X_1$ demand curves across all consumers. One point on the aggregated $X = X_1$ demand curve is relevant; identify it.
   - **c.** Draw a second demand curve that would characterize the individual's demand propensities when the aggregate market quantity of $X$ being consumed is $X^{MKT} = X_2 > X_1$.
   - **d.** Aggregate the individual $X^{MKT} = X_2$ demand curves. One point on the aggregated demand curve is relevant; identify it.
   - **e.** The market's demand curve for $X$ is the locus of points such as the two points that you have identified in response to questions b and d above. Now draw in the market's demand curve for $X$.
7. Return to the question raised in point 4: Does the bandwagon effect change the elasticity of the market demand curve for $X$? Did your analysis for point 6 confirm your intuitive response for point 4? If not, which do you agree with: analysis or intuition?
8. Repeat the steps outlined in point 6 for the snob effect and product $Y$.
9. Address point 7 as it applies to product $Y$.
10. Could either the bandwagon or snob effect result in the aggregate market demand curve for $X$ being positively inclined?

## TraderEx SIMULATION: CALL MARKET TRADING

The simulation software in *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course* can be used to demonstrate price and quantity determination in a market that comprises a large number of traders. In the context of a trading environment that is referred to as a *call auction*, we can see how buy and sell orders are integrated together to determine a common market clearing price, and exactly which orders do and do not execute. Call auctions are being widely used today in equity markets around the world to open markets at the start of each trading day, and to close markets at the end of each trading day. We will briefly introduce you to the calls here and then return to them in Chapter 4.

With call auction trading, orders are batched together for simultaneous execution in a big multilateral trade, at a single point in time, at a single market-clearing price. You can think of the market-clearing price as the equilibrium value ($P^*$) that we have modeled in this chapter. To obtain the clearing price, buy orders are sorted and cumulated, starting with the highest buy limit price and moving down to successively lower values, while sell orders are sorted and cumulated starting with the lowest sell limit price and moving up to successively higher values. The two cumulated curves are matched against each other, and the price that maximizes the number of shares that execute is set as the clearing price. The cumulated buy and sell curves look very much like the demand and supply curves that we have been dealing with, and the value that maximizes the number of shares that executes is the price where the two curves cross. Put it all together and perhaps you will agree that the call auction is a microeconomist's dream.

With this as background, you can turn to the TraderEx simulation. On the left of your TraderEx screen, you will see a book of orders that have already been submitted. Note the following:

1. Price is a discrete variable (we have set the minimum tick size at 5¢).
2. The buy orders and the sell orders are ranked. In accordance with the cumulation procedure we have just described, the TraderEx screen shows at each price the total number of shares that have been entered at that price and at all prices that are more aggressive (i.e., at higher prices for the buys and at lower prices for the sells). These cumulated numbers are buy and sell *schedules*.
3. The screen shows that shares are being offered for sale at prices that are equal to and lower than some of the prices at which shares are being sought for purchase. Consequently, trades will be made when the market is "called."

(*Continued*)

(*Continued*)

On the right of your TraderEx screen you will see graphs of the buy and sell schedules. Prices are on the vertical axis, and the cumulative number of shares to buy and to sell is on the horizontal axis. The graphs look like continuous curves but they are not. They have been created by plotting the price and size values from the buy and sell schedules and connecting the plotted points with straight lines. One other point is also shown on the graph: the point that identifies the market clearing price and the number of shares that will trade, as determined with reference to the buy and sell schedules. This single point is identified by dashed horizontal and vertical lines that extend to the axis. Notice that this point generally does not coincide exactly with the point where the graph of the cumulated buy orders crosses the graph of the cumulated sell orders.

In this simulation, we are giving you a last mover advantage. You can advance the simulation and see some more machine driven orders come in. Then the process will stop. When it does, you can enter your orders and, in so doing, affect the market clearing price. Of course you would not get this special treatment in an actual call auction, but for our purposes it is a useful device for showing a number of things.

This is what we would like you to do:

1. Before you enter any orders of your own, look at the cumulated buy and sell schedules shown at the left of your TraderEx screen and identify the clearing price that will maximize the number of machine entered shares that will trade. As you do this, keep in mind the following. At each price, the number of shares that will trade is the minimum of two numbers: the number of shares sought for purchase and the number of shares offered for sale. The clearing price that you are looking for gives you the maximum of these minimums.
2. With price being a discrete variable, we generally do not have a clean cross, which means that the number of shares to buy does not exactly equal the number of shares to sell. This being the case, a secondary rule of order execution is needed. In TraderEx (and in most call markets), time priority is the secondary rule (orders that have been entered first execute first). Noting this, and with the clearing price identified, specify which orders will execute and which orders will not execute at the call.
3. With reference to the graph of the buy and sell curves, we have noted that the single point, which is identified by dashed horizontal and vertical lines that extend to the axis generally does not coincide exactly with the point where the graph of the cumulated buy orders crosses the graph of the cumulated sell orders. Can you explain why there is a

discrepancy? (Obviously, the marketplace reality of price and quantity being discrete rather than continuous variables has its effect).

4. Determine the maximum number of buy orders that can be entered at the clearing price or higher without raising the clearing price.
5. Determine the maximum number of sell orders that can be entered at the clearing price or lower without lowering the clearing price.
6. The cumulated buy column and the cumulated sell column can be graphed. At this point, we will simply refer to the cumulated columns as "buy and sell curves." You can effectively shift the buy curve (while keeping the sell curve constant) by entering sufficiently large buy orders. Do so, and see what happens to the clearing price. Then compute the arc elasticity of the sell curve as the percentage change in transaction volume divided by the percentage change in price.*
7. You can effectively shift the sell curve (while keeping the buy curve constant) by entering a sufficiently large sell order. Do so, and see what happens to the clearing price. Then compute the arc elasticity of the buy curve as the percentage change in transaction volume divided by the percentage change in price.
8. Following up on questions 5 and 6, which is greater, the size of your order or the increase in the number of shares that trade? What explains the difference? For question 5, the answer depends on the elasticity of the sell curve; for question 6, the answer depends on the elasticity of the buy curve. Do you see it?
9. Verify that the less elastic the sell curve is in the neighborhood of the clearing price, the more the clearing price changes in response to the large buy order that you have entered at a high price. Similarly verify that the less elastic the buy curve is in the neighborhood of the clearing price, the more the clearing price changes in response to the large sell order that you have entered at a low price.
10. Large buy and sell orders placed in the neighborhood of the clearing price make these portions of the buy and sell curves more price-elastic; they also make the market more liquid. Do you see that elasticity and liquidity are essentially the same?
11. Enter orders into the call to verify that the buy and sell curves are more elastic when more orders are placed in the neighborhood of

---

*The term *arc elasticity* is used when the elasticity is measured using discrete price and quantity changes. With infinitesimal changes, the term "point elasticity" is used. To compute the percentage changes of price and quantity, for both price and quantity sum their initial values and their new values and divide by two.

(*Continued*)

(*Continued* )

the equilibrium price, and that they are less elastic when more buy orders are entered at higher prices and more sell orders are entered at lower prices.

12. If some participants are very bullish about a stock, they would be expected to place higher-priced buy orders in the call. Similarly, if some other participants are very bearish about the stock, they would be expected to place lower-priced sell order in the call. How would a greater divergence of expectations between the bulls and the bears affect the elasticity of the buy and sell curves in the neighborhood of the market clearing price?
13. Envision a sequence of calls, and consider the volatility of the clearing price from call to call. How might this volatility be related to the divergence of expectations between the bulls and the bears?

# CHAPTER 4

# Microeconomic Analysis Goes to Market

In this chapter we enter the nonfrictionless world. We all operate in real-world markets, and in Chapters 2 and 3 we have considered the theoretical properties of a frictionless environment. Our task now is to begin building a bridge between the two. The starting point for that bridge is not the reality of an actual market, but theory.

We first recognize that with adjustment costs, short-run responses to changing economic variables are only partial and, therefore, that demand and supply curves are less elastic in the short run than in the long run. After documenting an array of adjustment costs, we turn to a market attribute that is not germane to the frictionless environment but that is of critical importance in a nonfrictionless setting: liquidity. We emphasize that the liquidity of a market depends, among other things, on the design of a market, and our discussion of liquidity (read "illiquidity") brings us to our next topic: market structure. Getting key market structure issues under our belt positions us for the chapter's final topic, which is trading. With the use of demand curves, we identify a monetary measure of the gains a participant can realize from trading, and we look at a strategic order placement decision that equity traders face: how to price their orders. You will see that traders' decisions depend on the structure of the market that they are operating within.

Let's back up for a moment. Chapter 3 took us from the static decisions of individual participants to complete individual demand curves and, from there, to a market demand curve. At that point, with reference to our undefined product, *X*, we matched a market demand curve with a market supply curve and, in so doing, determined the price of *X* and the quantity of *X* sold in the market. Using the demand and supply curves for *X*, we explained that, in the frictionless marketplace with complete certainty, price is a rationing device that enables effective demand to be matched accurately and efficiently with available supply.

We used demand-and-supply analysis to show that when a price ceiling is imposed below the efficient market value, demand exceeds available supply and, consequently, that some other device for rationing the scarce supply must be resorted to. Similarly, when a price floor is imposed above the efficient market value, supply exceeds available demand and, again, a nonprice rationing scheme must be resorted to. Two other ways of altering a free market solution were also considered: a sales tax and a subsidy. Again, the pricing and output variables are impacted with questionable results.

We asked, How important is price as a rationing variable? Answer: Price's importance depends on the elasticities of demand and supply. So, what does elasticity depend on? Answer: The availability of substitutes. Understanding the alternatives and the ways in which market participants find them and respond to them is a big part of what microeconomics is all about. If the substitutes are there and the elasticities are substantial, then price is a good control mechanism that guides markets to efficiently allocate society's scarce resources among its competing needs.

But adjustments are difficult to make in the nonfrictionless world. When the price of gas at the pump triples in short order, pain is felt widely throughout the economy. The important thing to a microeconomist is that the pain gets people to respond in a myriad of ways that involve finding and using substitutes: buy more fuel efficient cars, drive less (do not travel as much, use public transportation more, ride your bike, or walk), wear heavier sweaters in the winter, build more fuel-efficient houses, invest in alternative energy sources, and so on. What our previous analysis has pointed out is that interference in the free setting of prices (the signals that guide individual decisions) stifles the various responses and, ultimately, impedes the implementation of effective solutions to dislocations caused by price shocks.

The nonfrictionless world is more complex and also more intriguing than its frictionless counterpart. As you will see, attaining an equilibrium price (an objective called *price discovery*) becomes difficult, and executable trades do not get made when the other side to a trade cannot be found (an objective called *quantity discovery*). When various costs, impediments, imperfect information, and other blockages affect participants' behavior in the marketplace, how far can theory take us?

To answer this question, we focus on a specific market: the market for equity trading. As we note in Chapter 1, equity markets are of major importance to the macroeconomy: They are intricate, they are interesting to know about, and equity investments are of substantial importance to a wide spectrum of individuals who own equity shares directly and/or through mutual fund holdings and pension plans. Theory adds important insight into the operations of the equity markets. As we note in Chapter 1, the relevant field in Financial Economics is *Security Market Microstructure*. The microstructure literature has taken us quite a distance down the road toward understanding this micro market. But the surgeon general has a warning: Theory has its limits, it can take us only so far. Our goal is to push that envelope as far out as possible.

So, let's get on with it. However, before turning to the equity markets, we stay with product *X* for one section longer.

## SHORT-RUN DEMAND AND SUPPLY FOR PRODUCT *X*

In one way in particular, microeconomic analysis has historically recognized that a marketplace is not a frictionless trading environment. Economists have long distinguished between long-run and short-run relationships. We did not introduce this distinction when obtaining the demand curves in Chapter 2. In effect and without saying so at the time, we obtained a long-run relationship between the price of *X* and the quantity of *X* demanded. How so? For the following reasons:

- The consumer's utility function is a description of the decision maker's tastes. Tastes are assumed to be given: They are what they are. We assume them to be constant and totally independent of any of the other variables that we have recognized as being important (for instance, the price of *X*, the price of *Y*, and income). The utility function translates into a family of indifference curves that reflect perfectly the decision maker's willingness to substitute between goods.
- We were analyzing a riskless world characterized by complete information. The decision maker knows all the input variables (the full set of relevant prices, income, and tastes), along with the quality characteristics of *X*. If everything is known, perfect decisions can be made.
- A change in the price of *X* relative to the price of *Y* changes the slope of the *X*, *Y* budget constraint. Our comparative statics showed the equilibrium consumption baskets for a decision maker both before and after a relative price change. Say $P_X$ falls relative to $P_Y$. Following this change, after the dust has settled, the consumer moves to a new basket that, compared to the old, contains more *X* relative to *Y*.

  But how long does it take to make the adjustment? The change certainly is not made instantaneously. It takes time. What is the path that is followed as the change is being made? We did not even attempt to deal with this one. Because neither the time taken nor the path followed

> was addressed, the analysis had to be long-run. Our comparative statics has shown only (1) the optimal consumption bundle before a price shock has occurred, and (2) the new equilibrium that is attained after the dust has totally settled.

Because of friction in the adjustment process, short-run adjustments are partial. How does this translate into short-run demand elasticities? Quite simply, the percentage change of *X*, namely, *dX*/*X*, is smaller in the short run than in the long run, and thus the short-run demand curve for *X* is less elastic than *X*'s long-run demand curve. That is, the absolute value of $\frac{(dX/X)^S}{dP_X/P_X}$ is less than the absolute value of $\frac{(dX/X)^L}{dP_X/P_X}$, where the superscripts *S* and *L* denote short-run and long-run responses, respectively.

The accentuated *in*elasticity of short-run adjustments compounds a public policy question that can attend a price shock, such as the jump in gas prices that we mentioned in the introduction to this chapter. Substitutability is great, substitutability brings us to desired, long-run solutions following a painful price increase for a basic commodity such as gas. But substitutes are not as available in the short run. Adjustment processes can be complex, painful, and lengthy. It is difficult to wait for long-run solutions, it is all too tempting to try to ameliorate a short-run problem and speed up the adjustments (or circumvent the problem) by interfering in the natural workings of a market. Can public policy come up with effective short-run solutions that do not perturb efficient long-run outcomes?

Some questions are more easily posed than they are answered and, with regard to this one, we have certainly witnessed a lot of debate. Clearly, there is no ready answer that can be generalized; the particulars of any given situation have to be taken into account. But there is a deeper question: How much confidence do we have, and can we have, in the natural workings of a free market?

Exhibit 4.1 presents the demand and supply curves diagrammatically. Exhibit 4.1(a) illustrates the relationship between the short-run and long-run market demand curves labeled $D^S$ and $D^L$, respectively, and Exhibit 4.1 (b) illustrates the relationship between the short-run and long-run market supply curves labeled $S^S$ and $S^L$, respectively. Notice that both short-run curves are more steeply sloped than their long-run counterparts: As we move down the demand curves from left to right, $D^S$ cuts $D^L$ from above and, as we move up the supply curves from left to right, $S^S$ cuts $S^L$ from below. Because the two short-run curves are steeper than their long-run counterparts and given that in both Exhibit 4.1(a) and Exhibit 4.1(b) $P_0$ and $X_0$ have the same respective values at the points of intersection, the two

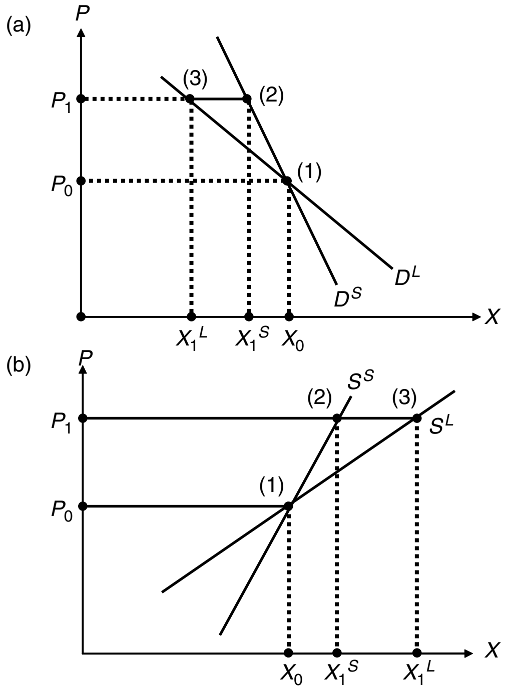


**EXHIBIT 4.1** Short-Run and Long-Run Demand [4.1(a)] and Supply [4.1(b)] Curves

short-run curves are clearly less elastic than their long-run counterparts at the points of intersection.

For both demand and supply, the effect of relative short-run elasticity on the quantity of $X$ demanded or supplied is also shown in Exhibit 4.1. Start at point (1) in both diagrams, where the $P_0$ and $X_0$ values represent a long-run equilibrium solution. Then let a shock occur that drives price in the short-run to the value $P_1$. A supply shift up and to the left could have sparked the price increase shown in Exhibit 4.1(a); or, alternatively, a demand shift up and to the right could have sparked the price increase shown in Exhibit 4.1(b). Notice that the short-run demand response in Exhibit 4.1 (a) results in the short-run equilibrium quantity of $X$ falling to only $X_1^S > X_1^L$, and that the short-run supply response in Exhibit 4.1(b) results in the short-run equilibrium quantity $X$ increasing to only $X_1^S < X_1^L$. These partial, short-run adjustments reflect the reality that substitution does not take place immediately and without cost.

## ADJUSTMENT COSTS

A myriad of factors can impede short-run adjustments. Somebody buys a new home and then a superior property becomes available at a good price; too late, that person thinks, I cannot face moving again so quickly. You have just bought a new suit and the price of a new suit falls; do you quickly run out and buy another one? Perhaps not, perhaps you simply decide to replace your new suit a bit sooner. The price of gas rises, but your gas-guzzling SUV is only six months old; do you sell it right away and buy a hybrid? Some people might, but others may not react so immediately. A firm builds a new plant and then experiences an unexpected increase in demand for its product; how long does it take for the firm to expand the size of its new plant or replace it with a larger one? Few firms will make the adjustment right away. A labor union signs a three-year contract with a company and then, in response to a demand shift, the free-market wage rate for its members rises; will the contract terms be renegotiated before the contract expires? Maybe not. A publicly traded company experiences a fifteen percent increase in earnings per share in the most recent quarter; will the company's management match this with a 15 percent increase in its quarterly dividends? A medical student is heading toward a career as a pediatrician when it becomes apparent that a shortage of radiologists is triggering price changes that favor this alternative field; will the student react to this news by switching fields? Some might if they are just starting their medical studies, but the information may have arrived too late for others who are further advanced in their programs.

Exactly what the specific adjustment costs may be differs from example to example, but they all have one thing in common: Their existence accounts for the delayed responses to a price change. How long does it take for an adjustment to be complete? That too differs from example to example. As we note in Chapter 1, the "long run" is an indefinite amount of time that depends on the example at hand. When it comes to labor contracts, plant size, and the supply of radiologists, the long run is indeed a lengthy stretch of time. With regard to other items such as suits and automobiles, the long run may well be shorter. Some adjustments in financial markets may take place relatively slowly (e.g., raising fresh equity capital) and others can occur with considerably greater frequency (e.g., a change of portfolio holdings of already-issued shares). Regardless, the economic impact of adjustment costs can be very significant, and the long run, short run distinction is critically important.

Let's hone in on a specific market to understand in greater detail how adjustment costs and the long run, short run distinction plays out. Let's turn our attention back to the secondary markets where already-issued equity shares trade. Equity trading is at the short end of the time spectrum—for equity trading, the time clock moves with extreme rapidity and the long run is not a very lengthy stretch of time at all. Who knows just how to measure it, but it is pretty much on the mark to say that, as of 9:30 A.M., at the opening bell, the long run for an equity trader is 12:00 noon.

With equities, the very act of trading is itself an adjustment process. This is because the demand to hold equity shares is stock-dimensioned, not flow-dimensioned. With flow-dimensioned demand, a consumer buys goods at a certain rate per unit of time (e.g., per day or per week) because he or she consumes the goods at that rate. For instance, if given the price of soda and all other related goods, you find it optimal to drink six cans of soda per week, then you must buy six cans of soda per week. The purchase of the six cans is equilibrium behavior. With a financial asset, you buy or sell shares only when you are out of equilibrium, only when you wish to hold more shares or fewer shares than are currently in your portfolio. With equities, trading frictions reduce your participation in the market, and affect your actions when you do participate. Let's look at these trading costs.

## Trading Costs

A spectrum of costs affects portfolio formation and trading. Here is a quick run-down on what is involved. Investment decisions are based on information, and information-gathering and -processing are costly. Trading (the implementation of an investment decision) is also costly. To understand just what the trading costs are requires knowledge of exactly how orders are handled and turned into transactions and transaction prices.

The costs of trading fall into two broad categories: *explicit costs* and *execution costs* (which, by their nature, are implicit). The explicit costs are visible and easily measured; they include, most prominently, commissions and taxes. Execution costs, on the other hand, are not easily measured; they exist because orders may, as a result of their size and/or the sparsity of counterpart orders on the market, execute at relatively high prices (if they are buy orders) or at relatively low prices (if they are sell orders).

Trading costs can appreciably reduce your returns. They also cause investors to adjust their portfolios less frequently and, accordingly, to hold portfolios that they would not deem to be optimal in a perfectly liquid, frictionless environment.

We facilitate our discussion of execution costs by defining the following:

**Quotation:** A price at which someone is willing to buy or to sell shares. A quote can be either "firm" or "indicative." If firm, the participant setting the quote is obliged to honor it if a counterparty arrives. If indicative, the quoting participant is not obliged.

**Bid quotation:** The price at which someone is willing to buy shares. The highest posted bid on the market is the "best market bid."

**Ask quotation (offer price):** The price at which someone is willing to sell shares. The lowest-posted ask on the market is the "best market ask."

**Limit order:** An individual participant's priced order to buy or to sell a specific number of shares of a stock. The limit price on a buy limit order specifies the highest (maximum) price a buyer is willing to pay, and the limit price on a sell limit order specifies the lowest (minimum) price a seller is willing to receive. Limit orders that are posted on a market are *prepositioned*. In other words, trading is a two-step process. First somebody prepositions (places) a limit order and waits for the arrival of a counterpart order to trade at market, at which time a trade occurs. The prepositioned orders to buy and to sell that are the most aggressive establish the best market quotes and thus the market's bid-ask spread.

**Market bid-ask spread:** The best (lowest) market ask minus the best (highest) market bid. The bids and offers can be market maker (dealer) quotes and/or the prices that individual participants put on their prepositioned limit orders. The market spread is sometimes referred to as the *inside spread* or as the *BBO* (best bid and offer).

**Market order:** A participant's unpriced order to buy or to sell a specific number of shares of a stock. A market order can execute against a dealer quote or against a prepositioned limit order. Market orders to buy are typically executed at the best (lowest) quoted ask, and market orders to sell are typically executed at the best (highest) quoted bid.

**Short selling:** A participant who does not own shares of a company but believes that the shares are overpriced can act on his or her opinion by *selling short*. A short sale involves selling borrowed shares. If price does fall, the short seller will buy the shares back in the market and return them to the lender. Short selling enables a participant to have a negative holding. Negative holdings are called *short positions*, and positive holdings are called *long positions*. A short sale that is executed without the shares having been previously borrowed is a *naked short*.

**Arbitrage trading:** Riskless arbitrage trading involves the coordinated buying and selling of the same asset so as to profit from the price spread that exists when the asset can be bought (in one market) at a price that is lower than it can be sold at (in another market). A *risk arbitrage* opportunity exists when a pricing discrepancy develops between two nonidentical assets whose returns are normally correlated with each other. For instance, one stock may be overpriced relative to another stock. We did not discuss it as such at the time, but arbitrage trading is a key element of the Capital Asset Pricing Model that we present in Chapter 3. For instance, if two stocks have the same covariance with the risky market portfolio, but one is trading at a lower price than the other, an arbitrageur will buy the underpriced stock (thereby pushing its price up) while simultaneously shorting the overpriced stock (thereby pushing its price down). And so, in a CAPM environment, it is arbitrage trading that ensures that all stocks are properly priced in relation to their covariance with the risky market portfolio.

Short selling facilitates arbitrage trading. Instead of buying cross-listed shares in market A and sending the shares to market B for sale, the arbitrageur simultaneously buys shares that he or she does not own in market A and sells shares that he or she does not own in market B. The sale in market B is a "short sale" because the sold shares are not owned. The arbitrageur then profits as the prices of the two stocks regain their proper alignment. We refer to arb operations in the next section of this chapter, where we focus on liquidity. As you will see, in the perfectly efficient, frictionless world of the CAPM, arbitrageurs are the ultimate source of the liquidity that characterizes that model.

Before proceeding, it is helpful to distinguish between active and passive traders. During normal trading hours, an execution is realized when two counterpart orders cross. This happens if one of the following occurs:

**Case 1:** A public trader first posts a limit order, and then another public trader submits a market order that executes against the limit order.

**Case 2:** A market maker (dealer) sets the quote, and then a public market order executes against the quote.

**Case 3:** Two or more public participants negotiate a trade. The negotiation may take place in a trading venue such as the New York Stock Exchange, the London Stock Exchange, Deutsche Börse, or

NASDAQ; in a brokerage firm in the so-called "upstairs market;" in an alternative trading system (quite a number of them exist today); or via direct contact between the two trading partners.

In each case, one party to the trade may be viewed as the active trader who is the instigator or liquidity taker, and the other party as the passive trader who is the liquidity maker. The one who is seeking to trade without delay is an active trader. Active traders are the public market order traders (cases 1 and 2) and the trader who initiates the negotiation process (case 3). Passive traders include the limit-order trader (case 1), the market maker (case 2), and the trader who does not initiate the negotiation process (case 3). If you are an active trader, you will generally incur execution costs; these payments are positive returns for passive traders. On the other hand, if you are a passive trader, you run the risk of a delayed execution or of not executing at all.

Armed with these definitions, we can identify the execution costs of trading:

**Bid-ask spread:** Because matched or crossed orders trigger transactions that eliminate the orders from the market, market bid-ask spreads are always positive and, with discrete prices, must be at least as large as the smallest allowable price variation (currently one cent in the United States for most traded stocks). An active trader typically buys at the offer and sells at the bid, and the bid-ask spread is the cost of taking a round trip (buying and then selling, or selling short and then buying). Conventionally, half of the spread is taken to be the execution cost of either a purchase or a sale (a one-way trip) by market order.

**Opportunity cost:** This term refers to the cost that may be incurred if the placement of an order (market or limit) is delayed (commonly in an attempt to achieve an execution at a better price), or if a trade is missed (because, for instance, a limit order has been placed but it does not execute). A buyer incurs an opportunity cost if a stock's price rises during the delay, and a seller incurs an opportunity cost if a stock's price falls during the delay.

**Market impact:** Market impact refers to the additional cost (over and above the spread) that a trader may incur to have a large order executed quickly. It is the higher price that must be paid for a large purchase, or the reduction in price that must be accepted for a large sale.

The bid-ask spread and opportunity costs are the major execution costs for a smaller, retail customer. Large institutional customers also incur

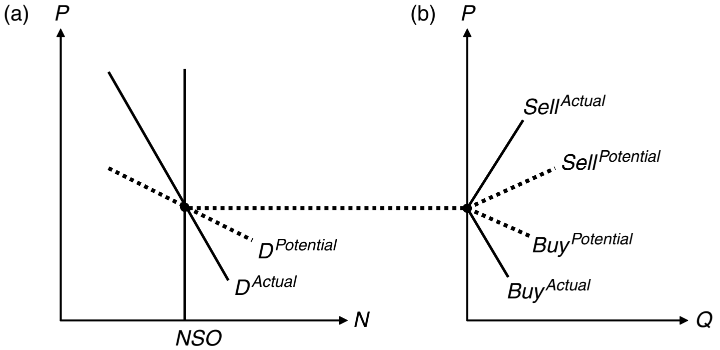


**EXHIBIT 4.2** Actual and Potential Demand [4.2(a)], and Buy and Sell [4.2(b)] Curves

market impact costs. None of these costs would exist in a frictionless market with no individual participant being large enough to have any impact on the price determined in the market. In the frictionless market, all portfolio desires would be fully and immediately expressed and satisfied. Essentially, a frictionless equity market would instantly achieve equilibrium. However, trading costs, because they reduce order submission and trading volume, can prevent equilibrium from being achieved. We have seen that adjustment costs reduce the price elasticity of both demand and supply for the unspecified good *X*. The same applies to the equity market: trading costs reduce trading volume. Effectively, both the explicit trading costs and the implicit execution costs result in the buy and sell order functions being less elastic. This reality is depicted graphically in Exhibit 4.2

Exhibit 4.2(a) displays two aggregate market demand curves, $D^{Potential}$ (which shows the number of shares that investors would like to hold at each price if trading were a frictionless process), and $D^{Actual}$ (which shows the number of shares that investors would like to hold when trading is costly). We have drawn the two curves so that they intersect each other at *NSO*, the number of shares outstanding. Notice that $D^{Potential}$ is more elastic than $D^{Actual}$. Exhibit 4.2(b) shows the market buy and sell curves that are associated with the two demand curves. Notice that $Sell^{Potential}$ is more elastic than $Sell^{Actual}$ and, similarly, that $Buy^{Potential}$ is more elastic than $Buy^{Actual}$. With regard to the order functions, the greater elasticities of the buy and sell curves in the absence of transaction costs show that transaction costs decrease the market's responsiveness to price changes.

Price elasticity is a term, however, that generally is not used much in equity trading. The term *liquidity* is employed instead. A liquid market has a lot of potential buyers who are standing by ready to respond to a price decrease and, in so doing, are there for sellers to sell to (if the price is right). A liquid market also has a lot of potential sellers who will respond to a price increase and, in so doing, will be there for buyers to buy from (also if the price is right). With good depth on both sides of a market, order imbalances have minimal effects on market prices, and relatively small price changes have a bigger impact on the number of shares that are traded.[1]

We have seen that the same applies for the undefined good, $X$. The short-run demand curve for $X$ is relatively inelastic because, in the short run, buyers respond only partially to price movements. If sellers were to bring additional goods to market, price would have to be lowered more to get buyers to purchase the extra amount. This *inelastic response* on the part of buyers can be thought of as equivalent to sellers finding the market to be relatively illiquid in the short run. Similarly, the short-run supply curve for $X$ is relatively inelastic because, in the short run, sellers respond only partially to price changes. If buyers are seeking to purchase additional units of $X$, the price of $X$ has to be raised more to get sellers to supply the extra amount. This *inelastic response* on the part of sellers can be thought of as equivalent to buyers finding the market to be relatively illiquid. Just as we have previously given much attention to elasticity, let us now focus on liquidity.

## LIQUIDITY[2]

"Liquidity" is a nebulous term that nobody can define with precision, but which most of us would probably accept is the ability to buy or to sell in reasonable quantity, in a reasonably short amount of time, and at a reasonable price (recall that we present this definition in Chapter 1). But what is reasonable? The inability to answer this question with precision suggests why in the finance literature formal liquidity models and empirical liquidity studies have not been forthcoming to anywhere near the extent that they have been for risk, the other major determinant of a stock's expected return. You might note that in our previous discussion of the risk and return trade-off and the demand to hold shares of a risky asset, we considered risk and return only, we ignored liquidity, and we got away with it because we were assuming a frictionless trading environment that, by being frictionless, is perfectly liquid.

---

[1] The depth and breadth of a market refers to the number of shares placed at any one price point (depth), and to the existence of orders at many different price points (breadth).

[2] Material in this section has been adapted from Marcus Hooper and Robert A. Schwartz, "The Limits of Liquidity," *Journal of Trading* (Summer 2008), 15–19.

## The Perfectly Liquid Environment

We have an excellent idea of what a perfectly liquid world might look like. In the frictionless environment of the Capital Asset Pricing Model (CAPM), an unlimited number of shares of an individual stock can be bought or sold immediately at a constant price. Peter Bernstein took us in this direction in his insightful article in the *Journal of Portfolio Management*.[3] In so doing, Bernstein underscored another way of viewing liquidity: In the completely liquid CAPM environment, the dominance of systematic risk over stock-specific risk with regard to asset pricing is complete, as we see in Chapter 3. But what does this dominance of systematic risk over stock-specific risk mean?

We see in the previous chapter that risk can be divided into two components, systematic risk and nonsystematic risk. We see that nonsystematic risk can be eliminated entirely by proper diversification, and that systematic risk gets priced. Moreover, two stocks that have the same systematic risk must be priced to yield the same expected return. As noted earlier in this chapter, arbitrage plays a key role in this CAPM result. Not only should all stocks (and portfolios) with the same systematic risk be priced to yield the same expected return, but the systematic risk of any stock (or portfolio) can be matched exactly by a proper portfolio combination of two or more other stocks or portfolios (as we discuss in Chapter 3). If any stock does not yield the same expected return as its matching stock or portfolio, an arbitrage opportunity exists: Short the stock with the lower expected return (it is overpriced), and acquire a long position in the stock with the higher expected return (it is underpriced).

In the frictionless environment, arb operations, being costless, instantaneous, and complete, immediately drive all prices into perfect alignment. That is why the frictionless world is perfectly liquid. That is why you can buy or sell unlimited quantities immediately at constant prices. Prices never get out of line; the arbs will always keep this from happening. In the frictionless world, arbitrageurs are the ultimate providers of unlimited liquidity.

One further condition is required for a market to be perfectly liquid, and it too is a property of the frictionless environment: Information collection, dissemination, and assessment must be costless and, therefore, instantaneous, for only then will all investors always be in complete agreement concerning the systematic risk and expected return of each and every asset. Perfect, riskless arb operations require this homogeneity of expectations. When this frictionless market condition is lost, different investors can form different expectations about systematic risk and expected return and, when they do, arbitrage will no longer be able to bring security prices into perfect CAPM alignment. We focus on imperfect information and divergent (i.e., nonhomogenous) expectations in Chapter 7.

---

[3] Peter L. Bernstein, "The Surprising Bond between CAPM and the Meaning of Liquidity," *Journal of Portfolio Management* (Fall 2007). Reprinted in *A Guide to Liquidity* (Institutional Investor, Fall 2007).

Alternatively stated, in real markets, which are replete with divergent expectations, equilibrium values are not immutable as they are under CAPM. Immutability requires that investors be in complete agreement about the risk and return properties of assets, and this is lost when expectations are divergent. While the standard CAPM models are based on the simplifying assumption that all investors have homogeneous expectations concerning the risk and expected return parameters for each and every asset in the market portfolio, the nonfrictionless world is characterized by divergent expectations.

### Back to Earth

For the real world, the concept of infinite liquidity is absurd. Illiquidity is the inevitable product of market frictions, and we clearly do not operate in a frictionless environment (and never will). No matter how efficient the trading mechanisms may be that are being used to handle customer orders, there are limits to the liquidity that participants can expect to receive. And the illiquid world is not a simple place to operate in. There are trading costs, both explicit and implicit; as a manifestation of trading costs, intraday price volatility gets elevated (especially at market openings and closings); and price discovery is considerably more difficult when different participants form different expectations about the risk and return parameters of different stocks. We address these manifestations of illiquidity in further detail in Chapter 7.

Implicit trading costs such as bid-ask spreads and market impact are manifestations of illiquidity, but they address only part of the story. To probe deeper, we focus on two major functions of a market: *price discovery* and *quantity discovery*.

1. **Price discovery:** The process of finding a value that best reflects the broad market's desire to hold shares of a stock is called *price discovery*. The root cause of price discovery being a complex process is friction with regard to the production, dissemination, assessment, and implementation of information. Information sets are huge, information bits are typically imprecise and incomplete, and our tools for assessing information are relatively crude. Consequently, different individuals in possession of the same publicly available information may naturally form different expectations about the risk and expected return parameters for any given stock or portfolio. And, when participants have different expectations, they

can also change their individual valuations at any time, including upon learning the valuations of others. This divergence of expectations and the attending interdependencies of different people's valuations make price discovery a complex, protracted process (we further examine the price discovery process in the application at the end of Chapter 7).

2. **Quantity discovery:** Large, institutional participants, because of their size and the size of their orders, typically approach the market wrapped in a veil of secrecy. This secrecy gives rise to a *quantity discovery* problem: How do the large participants find each other and trade? Quantity discovery may be viewed as liquidity discovery, and the incompleteness of quantity discovery as akin to latent liquidity (i.e., it is there if only it can be found). Large traders either "hold their orders in their pockets" until they are motivated to step forward and trade, or they send their orders to upstairs trading desks or to other nontransparent trading facilities, or they slice and dice their orders, feeding the pieces into the market over an extended period of time. One or more large buyers may be doing any or all of this at the same time that one or more large sellers are doing the same thing. Consequently, the contras may not meet, they may not be providing available liquidity to each other, and some trades that are in principal doable do not get made.

Equity trading, in short, is not a frictionless process.

## A Coordination Problem

Trading brings together the orders of many participants. We can think of trading as a process that coordinates orders that are being submitted by a spectrum of participants. Some people are seeking to buy when others are looking to sell. Some are patient while others want immediate executions. Some are large institutional customers and others are small retail customers. Some may be in possession of new information, some may have their own liquidity motives for trading, and others may be trading because their own share valuations differ from the market price of shares. And some might be trading for technical reasons because they believe that they have detected a pattern, perhaps a momentum move or a reversal situation.[4] As the orders from these different participants flow into the market, quotes are established, prices are set, and trades are made.

---

[4] Momentum and reversal behavior are noted in the bandwagon and snob effects application in the section titled "Let's Consider an Application" at the end of Chapter 3. How efficient are the trades? This depends on the efficiency with which the orders are coordinated as they are brought together and executed. At first thought this may seem simple, but it is not. Effective order coordination can be very difficult to achieve.

Proper order coordination is not simple for two reasons in particular: (1) participants have divergent expectations, as we discussed with regard to price discovery, and (2) participants differ markedly in size, as we discussed with regard to quantity discovery. Because price discovery under divergent expectations is a dynamic and imperfect process, participants may lack confidence in the share values that they see being established in the market. This lack of confidence leads them to employ trading strategies that may further undermine the sharpness of price discovery. And quantity discovery is an important and complex issue because large institutional orders, due to their size, incur market impact costs.

In an effort to control the implicit market impact cost of trading, institutional customers commonly break their large orders into little pieces so as to hide them in a river of similarly sized orders, or else they send them to nontransparent (dark) liquidity pools that, by being dark, conceal the presence of the orders that are in them. In the process, price discovery and quantity discovery have partly decoupled, thereby exacerbating the coordination problem. Price discovery largely takes place in a transparent exchange market that attracts much retail-order flow, while quantity discovery is increasingly taking place off-board in dark pools and other alternative trading systems. With the two functions inadequately coordinated, price discovery is destined to be inaccurate, and quantity discovery is bound to be incomplete. These, of course, are short-run effects. But in equity trading (as with most everything that we do), the short run is always with us.

Orders should ideally be handled in a way that leads participants to disclose their true desires to trade, regardless of how large or how small a customer might be. Doing so involves encouraging traders to submit their orders at aggressive prices, in a timely fashion, and in full size. Only when these three attributes of an order (price, time, and size) are fully represented in the market will liquidity be maximized. Designing a market structure that will achieve this goal is an ongoing objective of the market architects.

## Market Structure Matters

The translation of orders into transaction prices depends on the rules of order handling and trading that address the coordination problem and define the architecture of a marketplace. Much change and development has been observed on the market structure front in recent years for the equity markets. Nevertheless, despite striking technological advances, the emergence of new trading facilities, and a sharp intensification of intermarket competition in recent years, major problems persist, and equity market design remains a work in progress.

Two problems are paramount: (1) providing reasonable liquidity for mid-cap and small-cap stocks, and (2) amassing sufficient liquidity for large, institutional-sized orders for all stocks. With regard to both problems, we face a public policy conundrum: how to allow orders to flow freely into properly consolidated and appropriately transparent liquidity pools while at the same time ensuring that strong intermarket competition keeps driving our markets to ever greater efficiency.

At the heart of the conundrum lies the following reality. We are dealing with two markets, not one. First, there is the marketplace for the stocks that are traded, the arena where orders compete against other orders; this marketplace is most competitive when all customer orders are fully consolidated in one trading venue. Second, there is competition between different trading venues (a market for markets); this intermarket competition is strongest when there are multiple, alternative venues that orders can be sent to. Thus a trade-off exists: Increasing competition in one of these two markets (perhaps competition within the order flow) weakens competition in the other market (perhaps competition between the alternative trading venues). How should the trade-off be resolved? The issue has been debated for decades.

On a more fundamental level, much remains that is still not known about the dynamic processes of price and quantity discovery in an equity market. How can different trading modalities best be brought together in hybrid structures? In what new ways can technology be used? How dominant will the major exchanges remain vis-à-vis a variety of alternative trading systems, including banks and brokerage houses operating individually or in consortiums? The issues are complex, and the answers are diverse. Logical analysis can take us only so far. In the opinion of a free market economist, it is ultimately the market that will decide how to amass liquidity most effectively. Many of us buy into this opinion.

## The Big Picture

We live and operate in a nonfrictionless world. In that world, market structure matters because liquidity provision is complex. The complexity is in no small part attributable to trading involving price discovery and quantity discovery, to price discovery being imperfect, and to quantity discovery being incomplete. Underlying the complexities of both price and quantity discovery is the reality that investor expectations are divergent. The root cause of divergent expectations is the magnitude, complexity, imprecision, and incompleteness of information. And so, market structure ultimately matters because of the cost of receiving, processing, and acting on huge and enormously complex information sets. Perhaps with better technology and superior market structure the liquidity envelope can be pushed back. Nevertheless, we will never achieve a virtually frictionless environment. We will always be banging against the limits of liquidity.

## MARKET STRUCTURE

The forces of demand and supply that drive prices and quantities to equilibrium values are strong, and Adam Smith's invisible hand is a cool concept. Nevertheless, a real-world market needs structure. Where and how do participants meet? How are the prices that participants see literally set, revealed, and responded to? Deep down, what does a market really look like? A wide spectrum of answers to these questions exists. Some markets are auctions: people pack into a room and, under the direction of an auctioneer, call out or respond to alternative prices until the highest bid or the lowest offer is found. Some markets are trading posts where vendors and customers bargain (some say "haggle") over price. Many stores have *administered prices* (namely, some human decision maker sets them), and the vendors are not willing to bargain. Many markets are run by human agents; some are electronic (eBay, for instance). Some markets have direct access to vendors (stores typically operate this way) while others include middlemen; the intermediaries can be brokers (who handle your orders for you) or dealers (who, like the used-car dealer, you buy from or sell to). Some markets are *periodic* (like the auctions for art or cattle), many are continuous (you can buy or sell whenever the store is open), and some never close (like the 24-hour grocery store).

Considerable variety can exist within each of the broad categories. You might yourself think of various different kinds of markets that you have seen and dealt in. You might contrast, from your own experience, an art auction with the electronic purchase of airline tickets You might think of all the differences between the apartment/housing market and late-night shopping at a 24-hour supermarket, or compare bargaining for a better price with a new- or used-car dealer to buying or selling stocks in a financial market. Each of these micro markets has its own characteristics; each has its own distinctive ways of handling participant orders and turning them into trades.

In all of the micro markets that we have just mentioned, the pressure of Adam Smith's invisible hand can be felt. Nevertheless, does anyone actually see the underlying demand and supply curves that exert their pressure on price and output determination? Perhaps an econometrician has estimated the curves using data gathered from the past. But just how visible are these curves to those who operate in the markets, including the people who set the administered prices? Let's turn to the equity markets for already-issued shares and find out more about what might or might not be known in a micro market. We will deal with four generic structures:

1. Quote-driven (dealer) market
2. Continuous-order-driven market
3. Call auction
4. Negotiated (block) market

## A Quote-Driven (Dealer) Market

For a trade to be made in an equity market (or most anywhere else, for that matter), a price first has to be established by somebody. In a dealer market, dealers state the prices. That is, a dealer posts *two-sided quotes*: a bid quote ($P^{BID}$, the price at which the dealer will buy shares from you if you are looking to sell), and an ask quote ($P^{ASK}$, the price at which the dealer will sell shares to you if you are looking to buy). Because the dealer quotes "prime the market," a dealer market is commonly referred to as a *quote-driven market*.

Dealers are *intermediaries*—they stand between the investors who are looking to buy shares and the investors who are looking to sell shares. More specifically, dealers are *sell-side intermediaries*. "Sell-side" means that, in the market for intermediary services, they are sellers of the service. To whom are they selling the service? Their customers are the *buy-side* players, both retail and institutional, who need help when buying shares for, or selling shares from, their investment portfolios.

As we have noted, the two basic types of intermediaries in the financial markets are (1) dealers and (2) brokers. The difference between the two is that dealers participate by trading *with* their customers (a dealer sells to a public buyer from its own inventory, and buys from a public seller for its own inventory), while brokers simply handle orders as agents for their customers. In the financial markets, dealers who have the formal obligation to continuously post two-sided quotes (i.e., both bids and offers) are referred to as *market makers*. These are the people who we have in mind as we pursue the discussion, and we use the terms "dealer" and "market maker" interchangeably.[5]

Picture yourself in the following position. Suppose that you are a market maker in *xyz* shares and that the market is about to open. Assume (for simplicity, of course) that you are the only market maker (in other words, that you have a monopoly position when it comes to trading *xyz* shares). And make one further simplification: you always operate with a fixed spread (i.e., your $P^{ASK}$ minus your $P^{BID}$ is a constant). This means that your only decision variable is the price level at which to place your quotes. So, how do you set the level of your quotes?

---

[5] Intermediaries such as exchange market makers operate in a *dual capacity* by acing as dealers for some orders and as brokers for other orders.

Before a market opens, the market maker will take into account the price that *xyz* closed at in the previous trading session. The intermediary will also consider all relevant news (for the stock, for the industry that *xyz* is in, and for the broad economy) that came in over the news wire since the close of the previous trading session. The dealer will further check how the broad market is expected to open, will touch base with other traders, customers and whomever else might have a reasonable opinion, will float up some tentative prices and assess the response, and then give it his or her best shot. The objective, of course, is to find a level for $P^{ASK}$ and $P^{BID}$ that best balances the demand and supply curves that the dealer cannot see. What if the market maker gets it wrong?

In Exhibit 4.3 we show diagrammatically how the forces of supply and demand interact to guide (or perhaps we should say "push and pull") the market maker's quotes in the direction of the equilibrium price which is labeled $P^*$. Exhibit 4.3(a) depicts a situation where the market maker underestimates $P^*$, and Exhibit 4.3(b) depicts a situation where the market maker overestimates $P^*$. Notice that, in both diagrams, the variable on the horizontal axis is the *expected* arrival rate (the expected number of orders per hour), not the actual number of buy and sell orders that will be submitted.

This subtle change in the definition of the quantity variable (from actual to expected) reflects the reality that order flow is stochastic (i.e., random), and that all the market maker can do by properly setting the quotes is to balance the expected arrival rates over relatively brief periods of time (we arbitrarily took the interval to be one hour). Notice also that the quantity variable is *the number of orders* per hour, not the *number of shares* per hour. Yet it is the aggregate number of *shares sought for purchase* that must be balanced against the aggregate number of *shares offered for sale*. However, it is easier to speak in terms of orders, and we can do this by simply assuming that all orders are of the same size (for instance, one round lot, which is 100 shares).

Notice also that the buy and sell curves in the two diagrams are step functions, not continuously downward or upward sloping curves. This property reflects the reality of price being a discrete rather than a continuous variable (in the diagrams we have taken \$0.05 as the minimum pricing increment). Continuous curves are easier to deal with and step functions are a pain, but discontinuity is a reality that we might as well honor in this diagram.

We are now in a position to understand better how the nontransparent forces of supply and demand (or, more precisely stated, "selling" and "buying" forces) guide the market maker to set his or her bid and offer quotes so

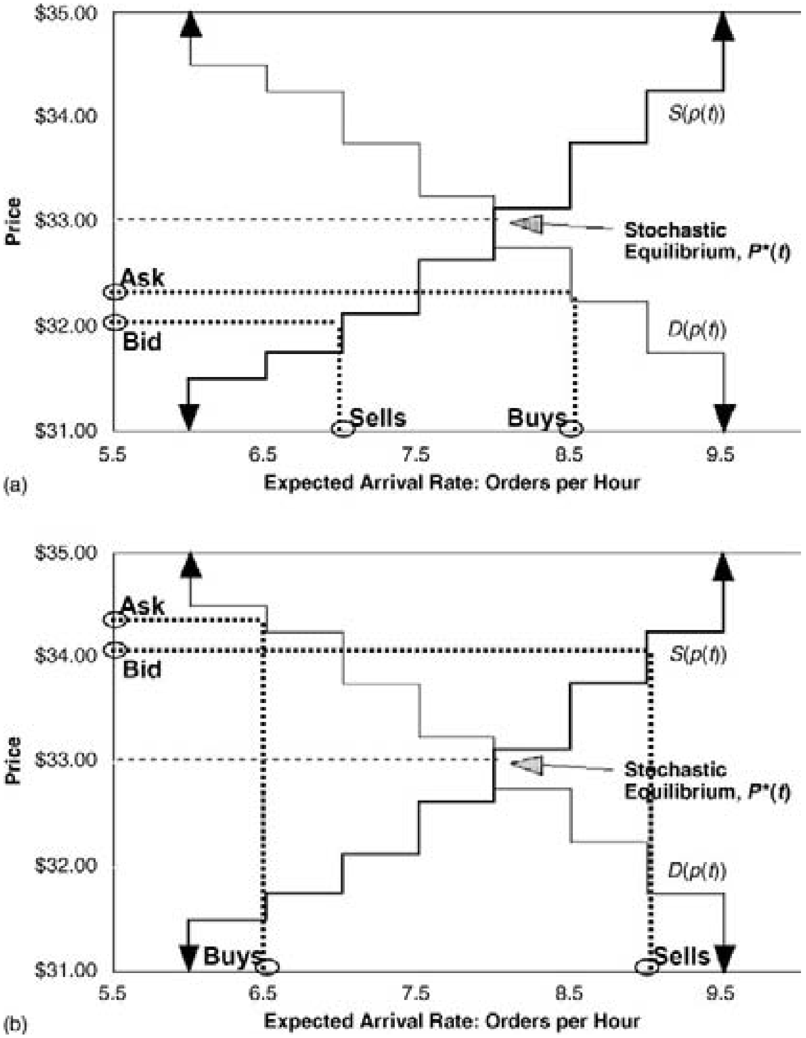


**EXHIBIT 4.3** Price Discovery in a Quote-Driven Market: Market Maker Quotes Are Too Low [4.3(a)], or Too High [4.3(b)]

that the quotes will tend to straddle the unobservable equilibrium price, $P^*$. The key to understanding how this works is to recognize that inaccurate price discovery is costly to a market maker. We can see this with reference to Exhibit 4.3 where $P^*$ is the price that best balances the expected arrival rate of buy and sell orders. Keep in mind that $P^*$ is not observable in the dealer market. Moreover, as we have said, the buy and sell curves are based not on actual orders but on the number of orders that are *expected to arrive* in a relatively brief, future interval of time (e.g., the next hour).

If the quotes straddle $P^*$, the arrival of buys is expected to balance the arrival of sells, and the dealer's inventory is most apt to stay in reasonable balance. However, ex post, the actual arrivals of buys and sells will likely not balance exactly because actual rates generally differ from expected rates by some random amount. But far more importantly, a serious inventory imbalance can quickly develop if the market maker misjudges the location of $P^*$ or if, unbeknownst to the dealer, a news event causes $P^*$ to jump (either up or down).

Exhibit 4.3(a) depicts a situation where the demand for the stock is *higher* than the market maker has anticipated, either because of mistaken judgment or because of the occurrence of a bullish news event that the dealer has not yet learned about. We see in Exhibit 4.3(a) that $P^*$ is 34.00 while the dealer's quotes are 32.95 bid, 33.05 offered. As buyers jump on the 33.05 offer, the rate of dealer sales to the public rises and the rate of dealer purchases from the public falls. With the order flow to the market maker out of balance, the dealer is soon faced with a rapidly increasing short inventory position. As an *inventory control mechanism*, the dealer raises the quotes, hoping once again to straddle $P^*$ and to return to a *flat inventory position* (i.e., be neither long nor short). The higher quotes work because the higher bid encourages public sales to the dealer, and the higher offer discourages public purchases from the dealer. In the process of adjusting the quotes, however, the market maker will be buying shares at a price that is higher than the price at which he or she had previously sold, which shows that the initial mispricing was indeed costly to the intermediary. But continued mispricing would be even more costly.

Exhibit 4.3(b) depicts the opposite situation: The demand for the stock is lower than the market maker has anticipated, either because of mistaken judgment or the sudden occurrence of bearish news that the dealer does not yet know about. Let $P^*$ be 32.00 while the dealer's quotes are 32.95 bid, 33.05 offered. As sellers jump on the 32.95 bid, sales to the dealer increase and purchases from the dealer fall. Now, the market maker is faced with a rapidly growing long inventory position. In response, the intermediary will lower the quotes, trying once again to straddle $P^*$ and return to a flat inventory position. In the process of adjusting the quotes, however, the market maker will be selling shares at a price that is lower than the price at which he or she had previously bought. Once again we see that the initial mispricing was indeed costly to the market maker.

Based on our consideration of Exhibit 4.3, we conclude that the market maker will indeed feel the pressure of the underlying buy and sell curves even though he or she does not see them splayed out as they are in Exhibit 4.3. The link that connects the market maker's pricing decisions with the unobservable equilibrium price is the order flow that the dealer receives based on his or her quotes. The order flow impacts the dealer's inventory of shares, and the intermediary resets the quotes in an attempt to steer his or her inventory back into balance (i.e., to be neither too long nor too short). The resetting of the quotes involves moving them in the direction of $P^*$, and the process continues until $P^*$ is straddled (unless, following the arrival of new information, $P^*$ jumps to a new value before it is straddled).

At this time you might look again at Exhibit 3.8 and the associated discussion of price setting with respect to the generic product, *X*. The earlier discussion was relatively abstract while the current one is relatively concrete. Nevertheless, the two stories concerning the forces of demand and supply are basically the same.

### A Continuous-Order-Driven Market

Dealer markets are the prevalent trading modality for many financial instruments such as fixed income, but they are not used today nearly as much as in the past for trading the equity shares of publicly listed companies (especially the big-cap companies). Instead, order-driven markets have become the preferred structure and, around the world, all of the major markets now offer electronic, order-driven trading. This is the market structure that we focus on in this subsection: electronic, limit-order book, order-driven markets.

We started our discussion of the quote-driven market by stating that, for a trade to be made, a price first has to be established by somebody. With an order-driven market, the limit orders placed by some investors who are looking to buy or to sell shares establish the prices at which other investors can sell or buy by market order. Because the public limit orders "prime the market," this market structure is referred to as an *order-driven market*.

An order-driven market can be a continuous trading environment (as is a quote-driven market) or a periodic-call auction, or a hybrid combination of continuous and periodic call-auction trading. A continuous market is "continuous" because time is continuous, not because trading is continuous. Trades are made discontinuously (i.e., sporadically) over time, but in continuous trading they can be made at any moment in time that the price of a buy order matches or crosses the price of a counterpart sell order. We focus on the continuous order-driven market in this subsection, and on the periodic-call auction in the subsection that follows.

Brokers are included in real-world order-driven markets to handle customer orders, and a dealer role can also be included (in fact, it has been included in many primarily order-driven markets) to bring further liquidity to the market, but we need not take these sell-side participants into account in this discussion. We simply assume the existence of a book that contains customer limit orders, and do not worry about how the orders are technically handled and entered into the book, either by sell-side agents or by buy-side participants having direct market access. What we care about right now is the economic functioning of the book itself. In the past, limit orders at the stock exchanges were literally written by hand and kept in a book. This practice ended with the introduction of electronic order management, but the term *book* continues to be widely used.

Now the big question: How do the forces of supply and demand guide the continuous order-driven market in the direction of an unobservable, underlying equilibrium price? We address this question with reference to Exhibit 4.4, which displays a limit-order book. The limit-order book in Exhibit 4.4 is populated by orders that have already come in to the market, not orders that are expected to arrive (as is the case in Exhibit 4.3 for the dealer market). Notice that the orders on the book have been entered at an array of different prices. The higher-priced orders to the right of the price column are limit orders to sell; they are referred to as *offers*. The lower-priced orders to the left of the price column are limit orders to buy; they are referred to as *bids*.

At each price point, we do not show the individual orders, but the total number of shares for all of the orders that have been placed at that specific price. With continuous trading, there are no buy orders on the book with prices that are equal to or higher than the price of any sell order (or any sell orders that are priced at or below any of the buy limits on the book). Rather, there is a clean separation of the buy and sell orders because any *crossing orders* (that is, buy and sell orders that "cross in price") would have executed against each other and, in so doing, would have been eliminated from the book. And so, typically, there is always a positive gap between the lowest offer and the highest bid. The gap may be as small as the minimum tick (or *price variation*, as it is also called), which in Exhibit 4.4 is 0.10, but often the gap is larger. The gap is the *bid-ask spread*.

And so, as in the dealer market, a bid-ask spread is established in an order-driven market. The spread is set as follows. The lowest of the offers (i.e., the price of the lowest priced sell limit order) is the ask quote for the market: $P^{ASK}$. The highest of the bids (i.e., the price of the highest-priced buy limit order) is the bid quote for the market: $P^{BID}$. In Exhibit 4.4, the best market quotes are 10.60 bid and 10.80 offered. There is one empty price point in the exhibit within the spread at 10.70, and either a bid or an offer could be entered at this price point without triggering a transaction.

**EXHIBIT 4.4** Limit-Order Book for a Continuous Market

| Bids | Price | Offers |
|---|---|---|
| | 11.60 | |
| | 11.50 | 29 |
| | 11.40 | 22 |
| | 11.30 | |
| | 11.20 | 5 |
| | 11.10 | 92 |
| | 11.00 | 9 |
| | 10.90 | |
| | 10.80 | 6 |
| | **10.70** | |
| 18 | 10.60 | |
| 19 | 10.50 | |
| | 10.40 | |
| 41 | 10.30 | |
| | 10.20 | |
| 42 | 10.10 | |
| 21 | 10.00 | |
| | 9.90 | |
| 44 | 9.80 | |
| 27 | 9.70 | |

Commonly, the spread will include one or more empty price points (i.e., be wider than the minimum tick size).

In the continuous order-driven market, transactions are triggered by the arrival of market orders that execute against the posted limit orders. We see in Exhibit 4.4 that the best-priced offer is six shares at 10.80.[6] A market order to buy would execute against one (or more) of the limit orders at this price.

With both market and limit orders in the picture, there are two ways in which you can interact with the order-driven market as a buyer:

1. You can attempt to purchase shares by entering a limit order at any price you choose that is lower than the best (lowest) posted offer (we say "attempt" because your limit order will execute only if a market order to sell arrives and executes against it, and this may not happen or it might take an appreciable amount of time to happen).

---

[6] The number of shares shown in Exhibit 4.4 are all very low, but fit in with our keeping things simple. Stocks are commonly traded in round lots of 100 shares and, if you wish, you can think of the entries in Exhibit 4.4 as the numbers of round lots.

2. You can enter a market order and buy with certainty at the offer.[7]

Similarly, if you want to sell shares, you can do so in two ways:

1. You can attempt to transact by posting a sell limit at any price you choose that is higher than the best (highest) posted bid.
2. You can enter a market order and sell with certainty at the bid.[8]

Market orders have one benefit over limit orders: They give you certainty of transacting (with a limit order you run the risk of not realizing an execution). Limit orders have one benefit over market orders if they execute: they do so at better prices. When your limit order executes, you have *saved the spread* (a market order to buy executes at the higher asking price, and a market order to sell executes at the lower bid price, while your limit order executes at the price at which you have placed it). You might look again at Exhibit 4.4 with this thought in mind to make sure that it makes good sense to you.

Prices change in an order-driven market as limit orders are entered into the book and as they leave the book. A limit order will leave the book if an arriving counterpart market order executes against it or if the investor who placed the order changes his or her mind and either rewrites the order or simply cancels it. Any of the above is particularly likely to happen after the advent of new information, but new information is not the only instigator of change. Any individual participant may reassess his or her own opinion about a stock, and/or reconsider his or her cash position if, for instance, a tuition bill has just arrived or a winning lottery ticket has just been bought. Technical traders might jump on board as momentum players or as contrarians (people who tend to buy when others are looking to sell, or vice versa). The important thing is that the individual demands to hold shares of a stock will shift and, as they do, the equilibrium price in the market will change.

How does the order-driven market respond to an order imbalance? Suppose that, for any of the above reasons, a preponderance of investors suddenly turn more bullish. While both buy and sell orders may continue to flow into the market, the following are most apt to occur:

---

[7] A price limit can also be put on a buy order at a value that is equal to or higher than the best offer. This order (called a *marketable limit order*) executes immediately (like a market order) but unlike a market order protects the participant from buying at a higher price in case the offer price rises while the buy order is being submitted.

[8] A price limit can also be put on a sell order at a value that is equal to or lower than the best offer. This *marketable limit order* executes immediately (like a market order) but also protects the participant from selling at a lower price in case the bid price falls while the sell order is being submitted.

1. Market orders to buy will come into the market and eliminate limit sells from the book.
2. Limit sell orders at the lower prices will be repriced upward or cancelled.
3. Limit orders to buy will be entered at higher prices and the buy side of the book will fill in at higher prices.

With these three adjustments, the best market bid and offer quotes will move up. On the other hand, if a preponderance of the investors suddenly turn more bearish:

1. Market orders to sell will come into the market and eliminate limit buys from the book.
2. Limit-buy orders at the higher prices will be repriced downward or cancelled.
3. Limit orders to sell will be entered at lower prices and the sell side of the book will fill in at lower prices.

With these three adjustments, the best market bid and offer quotes will move down.

With these adjustment processes in hand, we can now see more clearly how the underlying forces of supply and demand *tend to* drive the bid and the ask quotes in the continuous order-driven market in the direction of an unobservable equilibrium price. Notice that we said, “tend to”: As in all real-world, nonfrictionless markets, there are exceptions.

With the information contained in the limit-order book, we can construct buy and sell order functions that resemble those that we present in Chapter 3. Recall that the limit-order book in Exhibit 4.4 displays the total number of shares at each price (as we have said, the total is the sum of all of the shares of all of the limit orders that have been entered at each specific price). Cumulating these totals *across the prices* gives us the buy and sell order functions that we are looking for.

For the buy orders, we cumulate from the highest priced buy to the lowest; this is because the limit prices on the buy orders are the *maximum amounts* that the customers are willing to pay (e.g., if you enter a limit order to buy 100 shares at 32.15, you will willingly buy the shares at 32.15 or lower, but have stated that you do not want to buy at any price that is higher). For the sell orders, we cumulate from the lowest priced sells to the highest; this is because the limit prices on the sell orders are the *minimum* *amounts* that the sellers are willing to sell down to (e.g., if you enter a limit order to sell 100 shares at 32.15, you will willingly sell the shares at 32.15 or higher, but have stated that you do not want to sell at any price that is lower).

Exhibit 4.5 shows the buy and sell order functions for the order-driven market. In Exhibit 4.5(a) we show the order book from Exhibit 4.4 with the addition of two cumulated columns (one for the buys and one for the sells). The values in these columns are shown graphically in Exhibit 4.5(b); notice that these are step functions, as they are for the quote-driven market (as shown in Exhibit 4.3). Exhibit 4.5(c) is the same as Exhibit 3.15(a); it displays representative buy and sell order functions for a frictionless environment in equilibrium, after all participants have submitted their complete (and continuous) order functions, and after all crossing orders have been eliminated by the trades that have been made.

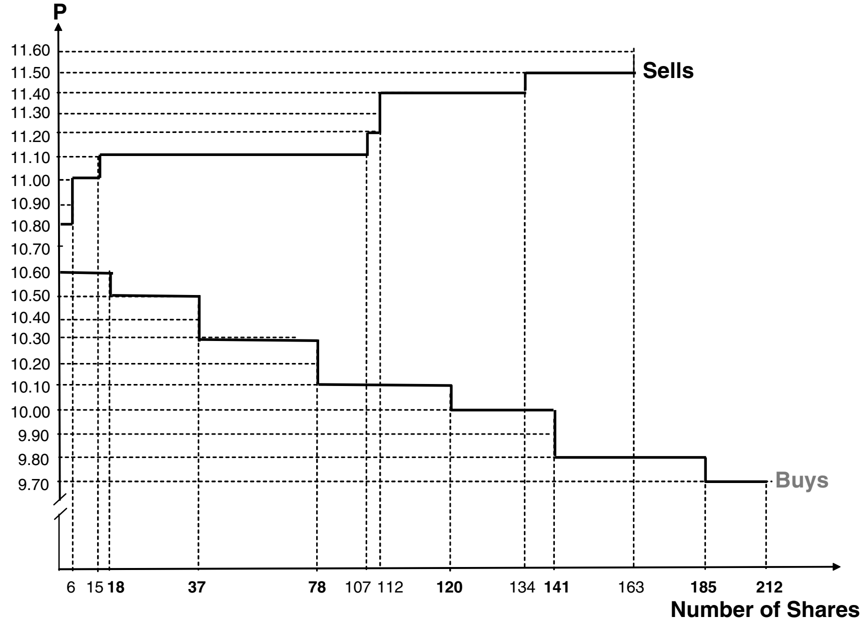
**EXHIBIT 4.5** Limit-Order Book for an Order-Driven Market: Cumulated Bids and Offers [4.5(a)], Their Graphical Presentation [4.5(b)], and a Stylized Graph of Market Buy and Sell Curves [4.5(c)]

(a)

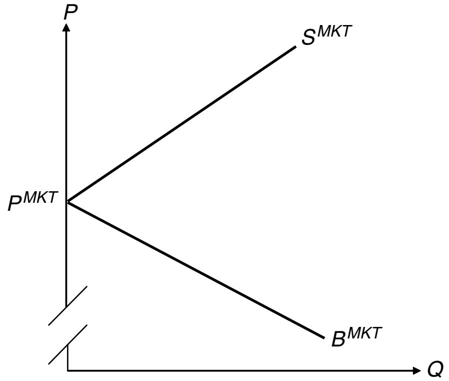
| Cum Bids | Bids | Price | Offers | Cum Offers |
|---|---|---|---|---|
| | | 11.60 | | 163 |
| | | 11.50 | 29 | 163 |
| | | 11.40 | 22 | 134 |
| | | 11.30 | | 112 |
| | | 11.20 | 5 | 112 |
| | | 11.10 | 92 | 107 |
| | | 11.00 | 9 | 15 |
| | | 10.90 | | 6 |
| | | 10.80 | 6 | 6 |
| | | **10.70** | | |
| 18 | 18 | 10.60 | | |
| 37 | 19 | 10.50 | | |
| 37 | | 10.40 | | |
| 78 | 41 | 10.30 | | |
| 78 | | 10.20 | | |
| 120 | 42 | 10.10 | | |
| 141 | 21 | 10.00 | | |
| 141 | | 9.90 | | |
| 185 | 44 | 9.80 | | |
| 212 | 27 | 9.70 | | |

(*Continued*)

(b)



(c)



Notice the similarities between Exhibits 4.5(b) and 4.5(c). In both, the buy order function is downward sloping and the sell order function is upward sloping. Both would translate into a demand curve to hold shares that is downward sloping. Both represent the forces of demand and supply. Notice also a difference. In Exhibit 4.5(b) there is a discrete bid-ask spread and the curves are step functions; in Exhibit 4.5(c) there is no bid-ask spread and the curves are continuous, linear functions.

There is another difference that is not apparent by visual inspection: Exhibit 4.5(b) is based on the subset of *customer orders* that have actually been sent to the market. Because trading is costly, not all potential investors participate in each trading session, and each individual order is only a partial representation of what an individual trader's complete demand curve might look like. Consequently, on both sides of the market, much of the desire to trade is not expressed, it is just latent. In contrast, Exhibit 4.5(c) is based on the complete individual demand curves of everybody who is in the market.

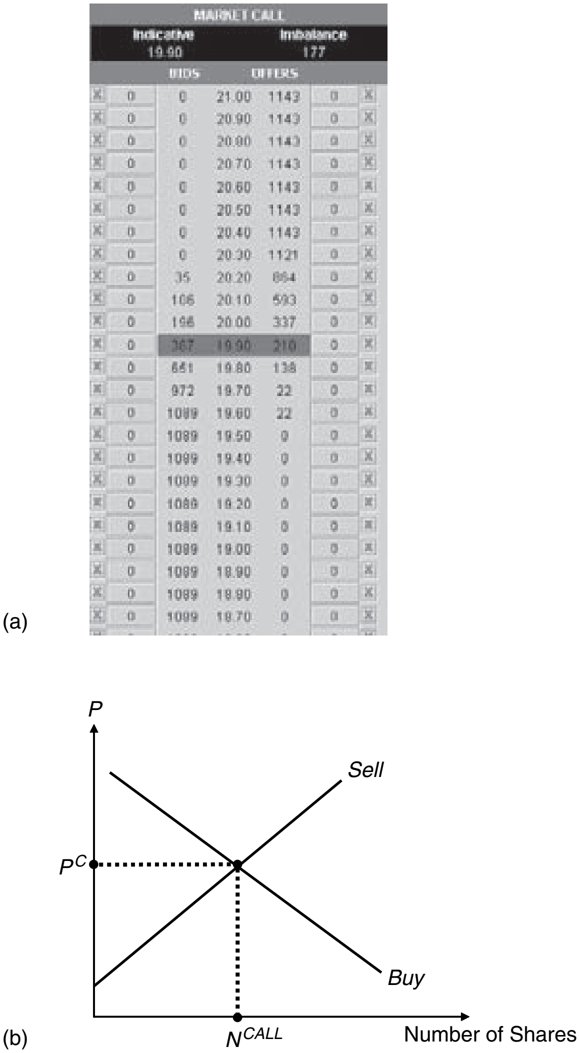
Notice that, for both 4.5(b) and 4.5(c), the more orders there are on the limit order book, the deeper is the book, and the flatter are the cumulated buy and sell order curves. A finance person would describe the deeper market as being more liquid; a microeconomist would say that the buy and sell curves are more elastic. As we have noted in the previous section of this chapter, liquidity and elasticity are very much two sides of the same coin.

Exhibit 4.5(c) is a theoretical construct while Exhibit 4.5(b) is real-world. Where does the equilibrium price come from, the theoretical construct or reality? Clearly, we want to apply the theoretical concept of an equilibrium price to an actual market. We want to know that economic forces are properly aligned so as to pull (or push) prices towards their equilibrium values. At any given moment, if the buy and sell orders that are flowing into the actual market are in roughly equal proportions, we can say that the order-driven market is in balance. But is it in equilibrium? This is more difficult to claim because at any given moment the buy and sell propensities are only partially expressed. Nevertheless, we can see how the natural market forces of demand and supply tend to push prices in the right direction (at least some of the time).

## A Call Auction

A call auction is an order-driven facility that provides periodic rather than continuous trading. With a call auction, counterpart buy and sell orders that would execute in continuous trading because they have matched or crossed in price do not transact immediately. Instead, all orders are batched together for simultaneous execution at a single price in a multilateral trade, at the specific point in time when the market is "called." Buy orders at the clearing price and higher execute as do sell orders at the clearing price and lower. Electronic call auctions are now being used by major stock exchanges around the globe both to open and to close their continuous trading sessions.

Exhibit 4.6 describes with an order book [4.6(a)] and with a graph [4.6 (b)] what a call looks like. You will see from the exhibit that the call auction



MARKET CALL

Indicative 19.90

Imbalance 177

|  |  | BIDS |  | OFFERS |  |  |
|---|---|---|---|---|---|---|
| ☒ | 0 | 0 | 21.00 | 1143 | 0 | ☒ |
| ☒ | 0 | 0 | 20.90 | 1143 | 0 | ☒ |
| ☒ | 0 | 0 | 20.80 | 1143 | 0 | ☒ |
| ☒ | 0 | 0 | 20.70 | 1143 | 0 | ☒ |
| ☒ | 0 | 0 | 20.60 | 1143 | 0 | ☒ |
| ☒ | 0 | 0 | 20.50 | 1143 | 0 | ☒ |
| ☒ | 0 | 0 | 20.40 | 1143 | 0 | ☒ |
| ☒ | 0 | 0 | 20.30 | 1121 | 0 | ☒ |
| ☒ | 0 | 35 | 20.20 | 864 | 0 | ☒ |
| ☒ | 0 | 106 | 20.10 | 593 | 0 | ☒ |
| ☒ | 0 | 166 | 20.00 | 337 | 0 | ☒ |
| ☒ | 0 | 387 | 19.90 | 210 | 0 | ☒ |
| ☒ | 0 | 651 | 19.80 | 138 | 0 | ☒ |
| ☒ | 0 | 972 | 19.70 | 22 | 0 | ☒ |
| ☒ | 0 | 1089 | 19.60 | 22 | 0 | ☒ |
| ☒ | 0 | 1089 | 19.50 | 0 | 0 | ☒ |
| ☒ | 0 | 1099 | 19.40 | 0 | 0 | ☒ |
| ☒ | 0 | 1099 | 19.30 | 0 | 0 | ☒ |
| ☒ | 0 | 1089 | 19.20 | 0 | 0 | ☒ |
| ☒ | 0 | 1089 | 19.10 | 0 | 0 | ☒ |
| ☒ | 0 | 1089 | 19.00 | 0 | 0 | ☒ |
| ☒ | 0 | 1099 | 18.90 | 0 | 0 | ☒ |
| ☒ | 0 | 1089 | 18.80 | 0 | 0 | ☒ |
| ☒ | 0 | 1099 | 18.70 | 0 | 0 | ☒ |

**EXHIBIT 4.6** Limit Order Book [4.6(a)] and Stylized Graph of Market Buy and Sell Curves [4.6(b)] for a Call Auction

sheds more light on the typically unseen forces of demand and supply than the continuous limit order book can do, and far more light than the quote-driven market is able to achieve.

Buy and sell orders are submitted to a call auction in a pre-call, order entry period. Some call facilities show the orders as the book builds, while others do not display the full book, for this is a market design feature that is not universally agreed upon by the market architects. Typically, however, at least an *indicated clearing price* is revealed to the public. This indicated price is what the clearing price would be if the call were held at that moment. It is a useful guide for participants who are deciding how to price and size their own orders.

Exhibit 4.6(b) is a stylized graphical picture of what the cumulated buy and sell orders might look like at the point in time when a market is called and the clearing price is established. On the graph, price is on the vertical axis and the cumulative number of shares sought for purchase or offered for sale is on the horizontal axis. To keep it simple, the lines shown in 4.6(b) are continuous (i.e., they are not step functions). With a continuous, upward-sloping sell curve, and a continuous, downward-sloping buy curve, a clean cross is guaranteed.

As shown in 4.6(b), the point where the two curves intersect establishes the market clearing price, $P^C$, and the number of shares that trade in the call, $N^{CALL}$. All shares sought for purchase that are priced at $P^C$ and higher execute, as do all shares offered for sale that are priced at $P^C$ and lower. Notice that, given the two cumulative order functions, the number of shares that execute when the market is called is maximized.

The maximization of trading volume is achieved precisely because the price set at the call, $P^C$, is determined by the point where the buy and the sell curves intersect. This key feature of call auction trading is consistent with what we have previously seen in Chapter 3, namely that the quantity of $X$ exchanged in the market is maximized at the price where the demand for $X$ and the supply of $X$ cross. The resemblance between the role of the order book in call auction trading and theoretical demand and supply analysis is, indeed, substantial.

Exhibit 4.6(a) is a tabular representation of orders that have been submitted to a call auction. Notice how the buy orders (bids) are cumulated from the highest priced buys to the lowest, while the sell orders (offers) are cumulated from the lowest priced sells to the highest. Consistent with Exhibit 4.6(b), the clearing price is the value that maximizes the number of shares that execute.

This is how the clearing price is determined when the order entry period ends, the book is frozen, and $P^C$ is set. At each value of price, the number of shares that would trade is the minimum of the cumulative number of bids at that price and the cumulative number of offers. For instance, at a price of 19.80, there are 651 shares sought for purchase, while 138 shares are offered for sale; the smaller of the two, the 138 offered for sale, is what the transaction volume would be if 19.80 was picked as the clearing price. Or at a higher price of, for instance, 20.00, there are 196 shares sought for purchase, while 337 shares are offered for sale; the smaller of the two, the 196 sought for purchase, is what the transaction volume would be at a price of 20.00.

The clearing price is the value that maximizes the minimums at the various prices. In Exhibit 4.6(a), 19.90 is the clearing price, and the minimum transaction volume at this price is the 210 shares offered for sale (387 shares are sought for purchase at this price). One price tick up, the minimum number is the 196 (<210) shares sought for purchase and one price tick down, the minimum number is the 138 (<210) shares offered for sale. So there you have it, 210 shares at a price of 19.90 is the very largest trading volume that can be squeezed out of this market.

Because the buy and sell price, quantity relationships for a real-world call are step functions, we generally do not have a clean cross at the clearing price, as is reflected in Exhibit 4.6(a). This, too, is a reality of a nonfrictionless market, and it too has to be dealt with. The most common way to resolve the rationing problem that is created by a nonclean cross is to time-stamp the orders and execute them according to the sequence in which they have arrived at the market (that is, *time priority* is used as the rationing device). In the example shown in Exhibit 4.6(a), where the trading volume at the call is 210 shares at the 19.90 price, 177 shares to buy (the 387 shares sought for purchase at the price less the 210 shares offered for sale) do not execute. Which orders are excluded? The last to arrive are excluded. They are the lowest in the time-stamped ranking and consequently they do not execute. Time priority is equivalent to the accountant's term, FIFO ("first in first out," which in this case means that, among the orders that would execute on the basis of price priority, the first order placed is the first to execute).

How close can a call auction bring us to nirvana? The procedure has a lot of appeal to a microeconomist—in a number of ways, it is close to being an economist's dream. Most importantly, liquidity is amassed at a single point in time, the demand and supply curves are visible, and the rules for order execution and price determination maximize trading volume for the given order flow. Still, keep in mind that trading costs inhibit order submission to all markets, including call auctions. Just as we saw with respect to continuous trading, not all participants will enter their orders into a call, and those who do so will not submit complete order functions. Rather, they place just one or at most a couple of order points. Thus order entry at a call is incomplete, aggregate propensities to buy and to sell remain partially latent, and a true equilibrium value is rarely attained with precision. Nevertheless, the forces of demand and supply are in operation, albeit imperfectly. As with the continuous market, the call auction, based on the order flow directed toward it, does tend to head in the direction of an efficient outcome.

### The Negotiated (Block) Market

Trades can also be made by two parties meeting each other directly and negotiating a trade. This market structure is best suited to handling the large block orders of big institutional investors.[9] A negotiation can be between two floor traders handling customer orders, or between an institutional customer and an upstairs block positioner, or by two large traders meeting each other directly. A negotiation can be a person-to-person process (either face-to-face or via telephone), or it can take place via an electronic interface. The alternative trading facility named Liquidnet is a good example of an electronic negotiation environment: Two customers can find each other on the Liquidnet screen and negotiate on both price and size. Pipeline, another alternative trading system, also brings the orders of large customers together electronically and accomplishes a similar goal, but Pipeline does not include an explicit negotiation process.

Negotiations do not take place in a vacuum, however. It simply is not realistic for two contra-parties to price a transaction in isolation from the rest of the market. Instead, the negotiators see the prices that are being set elsewhere for the stock that they are looking to buy or to sell. They check the prices in a continuous order-driven market, and/or in a quote-driven market, and/or in a call auction that is operating in parallel to their negotiation. They pay attention to these prices, and they base their negotiations on this pricing information. Consequently, the broader forces of supply and demand continue to exert their pressure even when the trades are made away from the mainstream.

## THE GAINS FROM TRADING

Strategic order placement is generally not on our minds when we go to the grocery store to buy our butter, eggs, and beer. Who would ever think of filling their shopping basket optimally so that their indifference curve is tangent to their budget constraint? Most of us just look at the price tag (which you can think of as the store's limit order) and toss the item into the shopping basket (which you can think of as hitting the store's limit order with your market order to buy). Then again, there are sales—do we look for them, wait for them, and respond to them? Sure we do but beyond this, the only strategic decision most of us make in the supermarket is the selection of the checkout line that we think will be the fastest.

---

[9] A block is typically defined as an order of 10,000 shares or more.

Beyond the supermarket, however, strategic thinking is required when we go to many markets to trade. More complex decisions such as the purchase versus the lease of a car; the ownership or rental of a home, plant, equipment, or office; the securing of a fixed versus variable rate mortgage for a home; the selection of a defined contribution or a defined benefits pension plan; the issuance of debt versus equity capital; the submission of a market order or limit order to buy or to sell shares (and, if the latter, the price to put on the limit order); and so on. In this section of the chapter, we focus on how we might think strategically when going to a secondary equity market to trade.

We have seen in the preceding section how orders are integrated together and turned into trades in four different market environments: quote-driven (dealer market), continuous order-driven (limit order book market), periodic order-driven (call auction), and negotiation. Many of today's equity marketplaces are hybrid combinations of these four modalities. How should a participant choose between them when going to a market with his or her order in hand? Answering this one calls for strategic thinking.

In each of the four modalities, we saw how market processes, in keeping with Adam Smith's invisible hand, guide market prices in the direction of theoretically desirable, equilibrium values. But the processes are imperfect. Various frictions prevent full participation on the part of traders and, consequently, when coming to the market, the representative participant makes a *strategic trading decision*. To formalize a strategic trading decision, we must establish the goal of the trader. This involves quantifying the monetary gains from trading and then determining how the gains might be maximized.

## Two Demand Curves

For any resource, a monetary measure of the gains an individual receives from trading can be obtained from that individual's demand curve. Let's return to our generic resource, $X$, and see how this works. Exhibit 4.7 presents two demand curves (for an individual) for $X$. They both start at the same intercept, a, on the price axis, they are both linear (for simplicity), but they decrease at different rates. The steeper curve is labeled $D^O$, where the superscript "$O$" indicates "ordinary." The flatter curve is labeled $D^R$, where the superscript "$R$" indicates "reservation." The ordinary demand curve is the curve that we have thus far been working with; it shows the amount of $X$ that the participant would want to acquire at each alternative potential price of $X$. The reservation demand curve is new; we explain the information that it reflects later in this subsection. First we show how consumer surplus, our measure of the gains from trading, can be determined using the ordinary demand curve for $X$.

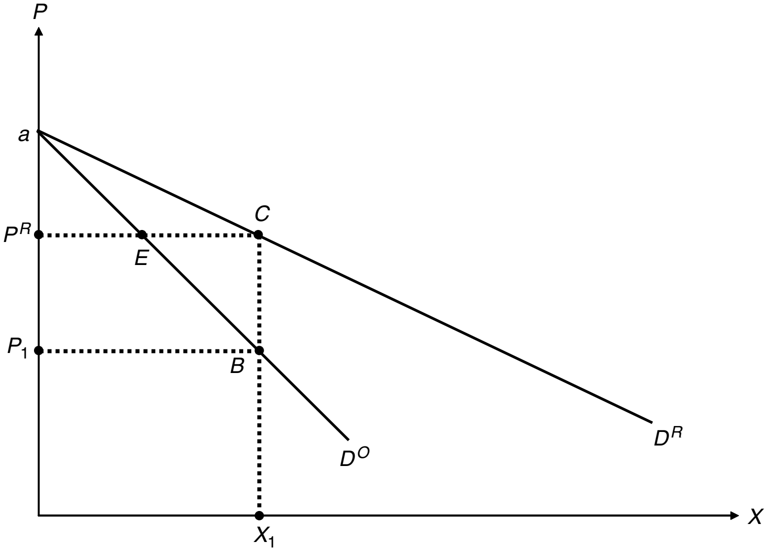


**EXHIBIT 4.7** An Individual's Ordinary Demand Curve ($D^O$) and Reservation Demand Curve ($D^R$) for a Generic Good *X*

## Consumer Surplus

The ordinary demand curve for *X* simply states that, at a price of, for instance, $P_1$, the consumer wants $X_1$ units of *X*. For the last unit obtained (the $X_1^{th}$ unit), the price of *X* reflects the marginal value to the consumer of the $X_1^{th}$ unit. If the price was slightly higher, than a bit less *X* would have been obtained and the marginal value from the last unit of *X* would have been a bit higher as well. Generalizing, each point along the ordinary demand curve shows just what the decision maker would pay, on the margin, for that last unit of *X*.

For instance, let *X*'s demand function be $X = 50 - 2P_X$. Using this function to assess discrete price and quantity increments, the numbers of units of *X* demanded are

| $P_X$ | $X$ |
|---|---|
| 24.5 | 1 |
| 24.0 | 2 |
| 23.5 | 3 |
| 23.0 | 4 |
| 22.5 | 5 |

and so on. So, if the price of $X$ is 23.5, the customer would obtain three units because the third unit is worth 23.5. A fourth unit is worth less, and it would be purchased only if the price was 23.0.

The first two units, on the other hand, are worth more than 23.5. At a price of 24.0, the decision maker would buy two units because the second unit is worth 24.0 (but the third unit is not worth this much). At a price of 24.5, the decision maker would buy one unit because the first unit is worth 24.5. But at a price of 23.5, the customer buys all three units and pays just 23.5 for each of them. So, here is the picture. At a price of 23.5, the customer would have been willing to pay:

24.5 – 23.5 = 1.00 *more* (if necessary) for the first unit, and
24.0 – 23.5 = .50 *more* (if necessary) for the second unit,

but, in the scenario that we are analyzing, these additional payments are not necessary. Restated slightly, the customer would have been willing to pay:

24.5 + 24.0 + 23.5 = 72.0 (if necessary) for the three units,
but in our scenario, he or she only pays 3 × 23.5 = 70.5.

So, he or she would have been willing to pay 72.0 – 70.5 = 1.50 *more* (if necessary) for the three units. This difference equals the sum of the two differences shown in the previous paragraph (namely, 1.00 + .50).

Now for the punch line:

> *The maximum differential amount that the customer would have been willing to pay if he or she* had to pay more *is a quantitative measure of the* gains to trade *that the customer realizes because he or she* does not have to pay more *(only 23.5 was paid for each of the three units). This measure of the gains to trade is called* consumer surplus.

The numeric illustration of consumer surplus is based on discrete numbers. Exhibit 4.7, which presents a continuous, downward sloping ordinary demand curve $D^O$, shows consumer surplus geometrically. In the graph, the consumer surplus associated with acquiring $X_1$ units all at the price $P_1$ is the area of the triangle, $aP_1B$. Here is why:

1. Start at the point labeled "$B$" in the exhibit. Point $B$ shows that at the price $P_1$, the decision maker obtains $X_1$ units of $X$.
2. Every price value along the $D^O$ demand curve from its vertical intercept at point a, down to point B, is the maximum that the decision maker

would be willing to pay for each marginal unit ranging from $X = 0$ to $X = X_1$.

3. Arbitrarily pick any point along the $D^O$ demand curve and drop a vertical line down to the horizontal line from $P_1$ to $B$. The length of that vertical line tells you how much more than $P_1$ the decision maker would have been willing to pay for that "last unit," if necessary.
4. In the exhibit, the area of the triangle reflects the sum of all of the vertical lines from $D^O$ to the horizontal line that extends from $P_1$ to $B$. The area, therefore, is the consumer surplus that is realized when all $X_1$ units are purchased at a per unit price of $P_1$. Thus, the area of the $aP_1B$ triangle is a monetary measure of the value, to our representative buyer, of obtaining all $X_1$ units at the same price of $P_1$. The area of this triangle, therefore, is the measure of consumer surplus.

### The Trader's Objective

As we have seen in Chapter 2, the trader's demand curve for $X$ is derived from the trader's utility function (which translates into a set of indifference curves) and budget constraint. The trader's ultimate objective is to maximize his or her utility subject to the budget constraint. The value of $X$ at each point on the individual's demand curve is the quantity of $X$ which, given the associated price of $X$, will in fact maximize the decision maker's utility function.

The maximum utility that the participant can achieve increases as the price of $X$ decreases and the decision maker moves down along his or her demand curve for $X$. Here is another way to say it: The lower is the price of $X$, the more $X$ the decision maker purchases, and the greater is his or her consumer surplus. Obviously, all else equal, the consumer is better off the lower is the price of $X$.

Commonly, the consumer simply observes a price and acts accordingly. There are times, however, when price is not simply responded to as a preestablished given. There are times when price can be negotiated, when price is not fully revealed ex ante, when price changes might be anticipated (for instance, sales and discounts). These are times when, in response to the uncertainties, a *trading decision* must be made. When making that decision, the participant can refer to a single objective that is consistent with the maximization of utility. That objective is the maximization of *expected consumer surplus*. As we will see in the subsection that follows, an investor, in placing an order to buy or to sell shares, can place that order so as to maximize his or her expected consumer surplus. In so doing, a *reservation price* is a very useful concept to have in hand, and to it we now turn.

## Reservation Price

Having identified consumer surplus, we can now introduce the term, *reservation price*. A reservation price is the maximum unit price that you would be willing to pay, if you had to, to get, for instance, $X_1$ units of $X$. We can identify this maximum with reference to Exhibit 4.7. In the graph, the reservation price for acquiring $X_1$ units is $P^R$ per unit of $X$, and $P^R$ is a point on the decision maker's *reservation price demand curve*, $D^R$. Here is why:

1. In Exhibit 4.7, Point $C$, with coordinates $P^R$ and $X_1$, lies on the reservation price demand curve, $D^R$ (to repeat, as a reservation price, $P^R$ is the highest unit price the individual would be willing to pay, if he or she had to, to acquire the $X_1$ units).
2. We have drawn the reservation demand curve so that its slope ($dP^R/dX$) is half the slope of the ordinary curve ($dP^O/dX$).
3. Because $(dP^R/dX) = \frac{1}{2}\, dP^O/dX$, the line segment from $P^R$ to E equals the line segment from $E$ to $C$. Consequently, the area of the triangle $aP^RE$ equals the area of the triangle, $EBC$ (with three angles and one side equal, the two triangles are identical).
4. Because the two triangles are identical, the rectangle $P^RCBP_1$ also measures the consumer surplus gained from consuming $X_1$ units all at the price $P_1$. [Notice that the consumer surplus triangle $aP_1B$ and the consumer surplus rectangle $P^RCBP_1$ have the area $P^REBP_1$ in common, and that, in changing from the consumer surplus triangle to the consumer surplus rectangle, we have subtracted one triangle ($aP^RE$) and replaced it with an identical triangle ($EBC$)].
5. When acquiring $X_1$ units at the price $P_1$, the individual's total expenditure on $X$ is $X_1P_1$. If, alternatively, the individual were to acquire $X_1$ units at the price $P^R$, his or her expenditure on $X$ would be $X_1P^R$. This higher expenditure would totally eliminate the individual's consumer surplus. At any price higher than $P^R$, the consumer would buy nothing, he or she would simply walk away from the trade. The reservation price for $X_1$ units is, therefore, $P^R$, and $P^R$ lies on the individual's reservation price demand curve, $D^R$.

You might now understand why in Exhibit 4.7 the slope of the reservation demand curve is half that of the ordinary demand curve [that is, why $(dP^R/dX) = \frac{1}{2}\, dP^O/dX$]. The reason is that only then are the two triangles, $aP^RE$ and $EBC$, identical, and only when they are identical does the area of the consumer surplus triangle $aP_1B$ equal the area of the rectangle $P^RCBP_1$, so that the rectangle can also be used as a measure of consumer surplus.

Our discussion has focused on a trader who is a buyer and whose gains to trade are measured by consumer surplus. We could similarly consider a seller whose gains to trade are called *producer surplus*. More generally, we can simply call the trading gains a *surplus*. For a seller, the slope of the reservation supply curve is half that of the ordinary supply curve (as long as the two curves are linear), and the seller's producer surplus from trading a certain quantity is given by both the triangular area above his or her ordinary supply curve for that quantity, and by the rectangular area established by the height of the ordinary supply curve at that quantity, minus the height of the reservation supply curve at that quantity, times that quantity. We suggest that you draw the picture yourself and verify this statement.

Going back to Exhibit 4.7 and the perspective of a buyer, you might wonder why anyone would actually obtain $X_1$ units at the price $P^R$ (which offers zero consumer surplus) when the smaller quantity associated with point E on the ordinary curve $D^O$ is an alternative that does offer consumer surplus. The answer is, if the two points *C* and *E* are both obtainable, the decision maker would indeed choose point *E*. For point *C* to be realized, the offer of $X_1$ units at the price $P^R$ would have to be an *all or nothing offer* that would make point E unobtainable (that is, your options are either to buy all $X_1$ units at a unit price of $P^R$, or to buy nothing at all).

## STRATEGIC ORDER PLACEMENT

With the terms *consumer surplus* and *reservation price* in our bag of tools, we can now focus on a participant's order placement strategy. As we do so in this section of the chapter, our attention is directed to the equity markets where shares of already-issued stocks are traded. We begin by making one simplifying and, indeed, very realistic assumption: that the trader has already made one decision: the number of shares ($Q$) that he or she wishes to buy or to sell.

### One Quantity: A Simplifying and Realistic Assumption

The "one quantity" simplification enables us to deal simply and directly with the price variable. Most assumptions that we make in the name of simplicity are unrealistic abstractions of the real world; the "one quantity" assumption is, in fact, quite realistic for three reasons in particular:

1. In some cases there is only one quantity. In an art auction, for instance, each painting or sculpture that is being auctioned is one of a kind. A home-buyer who has located his or her perfect abode will bid on that "one of a kind" house. When investing in stocks, many retail customers simply think in terms of buying or selling in, for instance, 100 or 200 share lots.
2. We can theorize about demand curves, but few of us know more than a couple of points on our demand curves. It is one thing to see a price and respond to it; it is another to say how the quantity you would demand would differ over a range of prices. This certainly is the case in securities trading. What participant would ever dream of calling a broker and saying, "I would like to buy shares of *xyz* corporation, my price intercept is 50 and my slope is –2"? Who would know these parameters, and how many brokers would understand what they were being told?

   The underlying economic problem is the size and complexity of information that underlies the values of the securities being traded. It is extremely difficult to assess very extensive and intricate information sets and to translate them into multiple price points that are functionally related to how many shares you would be looking to hold. Some large, sophisticated, generally institutional portfolio managers will do this and at times a multipriced, multisized order (referred to as a *scaled order*) is transmitted to the market. This, however, is rare. Generally, an order with one price and one size is submitted, and we are assuming that the size is exogenously given.
3. Submitting your complete demand curve to a continuous limit order book market would result in your loss of consumer surplus from trading. This is because, in such a market, prices are set while trading progresses. As we have seen for the continuous-order book market, a limit order sets the price for a trade that is made when the arrival of a market order triggers a transaction. When market orders clear out all of the limit orders at the most aggressive price on the book, they proceed to transact at the next price until it too is cleared out, and then move on to the next, and so on. Recognizing this, assume that you send the following set of buy orders to the market for a stock that you do not currently own, that the market ask is 10.10, that .05 is the minimum price variation (tick size), and that each of your orders reflects a point on your ordinary demand curve, $D^O$. Here is what your order would look like:[10]

---

[10] The continuous demand curve that underlies the schedule is $P = 10.10 - 1/500N$. Suggestion: Try verifying this by using the schedule to solve for the continuous-demand curve.

| Price | Total Number of Shares You Would Like to Own | Number of Shares in the Order Placed at Each Price |
|---|---|---|
| 10.10 | 0 | 0 |
| 10.05 | 25 | 25 |
| 10.00 | 50 | 25 |
| 9.95 | 75 | 25 |
| 9.90 | 100 | 25 |
| 9.85 | 125 | 25 |
| 9.80 | 150 | 25 |
| . | . | . |
| . | . | . |
| . | . | . |

Notice that the minimum tick size of .05 forces you to submit discrete (not continuous) points along your demand curve. What happens if one or more market sell orders arrive and price is pushed down until all of the buy orders through 9.80 are cleared off the book? Your first order executes at 10.05, your second order executes at 10.00, your third executes at 9.95, your fourth at 9.90, and so on. What happens to your consumer surplus? As the market marches down your demand curve, your consumer surplus is reduced (if the price and quantity adjustments were continuous your consumer surplus would be entirely eliminated). On the other hand, if you had submitted a limit order for 150 shares at 9.80, your expenditure for the 150 shares would have been 1,470.00 and your consumer surplus would have been 22.50 [= (10.10 – 9.80)150/2]. But, because you have scaled your order, your expenditure for the 150 shares is 1,488.75, which is 1.28 percent higher, and your consumer surplus is 3.75 [= (.05 × 25 × 6)/2], which is only 17 percent of what it could have been if a limit order for 150 shares at 9.80 were to execute. How do you protect yourself from this loss? Here is one answer: Do not submit your full demand curve, submit just one order.

Restricting quantity to a predetermined amount can be viewed as an extreme case of short-run demand inelasticity. Literally, we are assuming that the individual's demand curve has zero elasticity at all prices up to the price at which it has been placed. We have seen how adjustment costs reduce the elasticity of supply and demand curves. We now see how trading costs can reduce the elasticity of a trader's buy and sell order functions. So, how do traders price their orders if they prefix the sizes of their trading decisions?

## A Tale of Two Trading Venues

The answer to the question we just asked depends on the venue the traders are sending their orders to. We consider two venues in this subsection: a continuous order-driven market and a periodic call auction. We deal with each in turn and, as we do so, we assume that the one we are focusing on is the only available venue for the traders to use (that is, we do not consider a hybrid combination of continuous and call auction trading). To further simplify the discussion, we also assume commissionless trading.

**The Continuous Order-Driven Market** We can gain insight into order placement in a continuous order-driven market by considering a highly simplified environment. Assume that the investors who are trading shares in *xyz* corporation can be divided into two groups, one that is relatively bullish about the stock and another that is relatively bearish about the stock. In all other respects, let the members of each group be identical. The bulls are the potential buyers, and the bears are the potential sellers. Let 55 be the reservation buy price for the bullish investors, and let 45 be the reservation sell price for the bearish investors. The market is a continuous trading environment, and the participants arrive at it sequentially, one at a time. Upon arrival, each new participant first checks the limit order book to see what orders the earlier arrivals have placed, and then places his or her own order, which can be either a market or a limit order (but, regardless, the order is always of one fixed size). The trader's objective is to maximize the expected value of consumer surplus, and the maximization is achieved by proper order selection.

With the addition of a few more simplifying assumptions and a considerable amount of derivation, the equilibrium quotes that would be posted on the market can be determined.[11] Suffice it for us to say a few words about what this equilibrium would look like:

1. The bulls post limit orders to buy that are priced below their reservation buy price of 55. Thus as the limit orders to buy execute, the bulls gain consumer surplus.
2. The bears post limit orders to sell that are priced higher than their reservation sell price of 45. This way, as the limit orders to sell execute, the bears gain consumer surplus.
3. The limit buy orders of the bulls are priced lower than the limit sell orders of the bears. Accordingly, a positive bid-ask spread prevails in

---

[11] The formulation is presented in Puneet Handa, Robert Schwartz, and Ashish Tiwari, "Quote Setting and Price Formation in an Order-Driven Market," *Journal of Financial Markets* 6 (2003), 461–489.

the market, and market order buyers and market order sellers both realize positive consumer surplus from their trades.

4. The greater the percentage of participants who are bulls, the higher will be the location of the bid-ask spread in the market. With the quotes being higher and there being fewer sellers in the market, expected consumer surplus for the bulls is lower, while for the bears it is higher.
5. The greater the percentage of participants who are bears, the lower will be the location of the bid-ask spread in the market. With the quotes being lower and there being fewer buyers in the market, expected consumer surplus for the bears is lower, while for the bulls it is higher.

In this equilibrium, a participant can place either a market order or a limit order. Let's focus primarily on the latter because the market orders do not set the prices. A participant prices a limit order in relation to his/her expectation that a counterpart trader will arrive at the marketplace and, by placing a market order, trigger an execution. The limit orders that are placed establish a bid-ask spread. The location of the spread depends on the proportion of bulls versus bears, and it can be shown that the size of the spread is greatest when the proportion is fifty-fifty.

**The Periodic Call Auction** Order placement in a continuous order-driven market is relatively complicated. On the other hand, order placement in a call auction is relatively straightforward for a trader who is not large enough to be anything other than a "price taker." We explain what is involved with reference to Exhibit 4.8, which is a modification of Exhibit 4.7. While Exhibit 4.7 presents the investor's two *demand curves* to hold shares (the ordinary curve and the reservation curve), Exhibit 4.8 shows the *ordinary buy curve* ($B^O$) and the *reservation buy curve* ($B^R$) for a participant who is seeking to acquire more shares (the analysis is symmetrical for a participant who is looking to sell shares). If the decision maker currently is not holding any shares in his or her portfolio, the two buy curves in Exhibit 4.8 are comparable to the two demand curves in Exhibit 4.7.

In Exhibit 4.8, price is on the vertical axis and the size of the participant's order, $Q$, is on the horizontal axis. Because we have assumed that the decision maker has already decided the order size that he or she wants to work with, only one point on the horizontal axis matters. As shown in the exhibit, that order size is $Q_1$. With quantity set at $Q_1$, the only decision to make is how to price the order that will be sent to the call auction. In seeing how this decision is made, be sure to keep in mind the following property of call auction trading: As long as the price of a buy order is equal to or higher than the clearing price at the call (namely, $P^C$), that order will execute

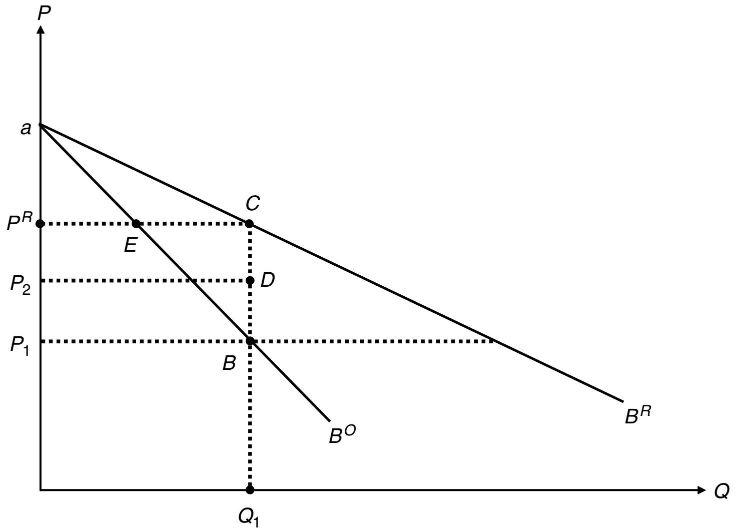


**EXHIBIT 4.8** An Investor's Ordinary Buy Curve ($B^O$) and Reservation Buy Curve ($B^R$) for Shares of a Risky Asset

(assuming a clean cross), *and* it will execute, not at its own limit price, but at the common clearing price that is set when the market is called.

Suppose that you are the trader who is submitting the buy order to the call, and that you are concerned only with buying shares today (in other words, assume that buying tomorrow is not a substitute for buying today). From your ordinary buy curve, we see that you would want to buy $Q_1$ shares at the price $P_1$. The price $P_1$, however, would not be a good price to put on your order. What if the clearing price, $P^C$, is a bit higher, say $P_2$ in the graph? Your limit order would not be priced high enough, and it would not execute. Yet, if it had executed, you would have realized consumer surplus in the amount of $P^RCDP_2$. Of course, if you had been able to acquire the $Q_1$ shares at $P_1$, your consumer surplus would have been the larger amount, $P^RCBP_1$, but that is beside the point. With a clearing price of $P_2$, your order would not have executed with an order price of $P_1$ on it, and your realized consumer surplus would have been zero.

Perhaps you should place your order at $P_2$? That would not be smart either: The clearing price could turn out to be a hair above $P_2$, and thus an order written at $P_2$ would not execute. Again, you would wind up with no consumer surplus even though you would have benefited from an execution at a clearing price in the neighborhood above $P_2$.

So, perhaps you should place your limit order at a yet higher price? Sure, you should. In fact, go all the way up to $P^R$ and see what you have. If the clearing price turns out to be $P^R$ exactly, your limit order executes, but you get no consumer surplus. However, you do not care, you have not lost anything either (remember, we have assumed commissionless trading). If the clearing price is greater than $P^R$, the limit order does not execute, but again you do not care, you do not want to pay more than your reservation price anyway. Now for the good part: If the clearing price is any amount below $P^R$ (it might even be lower than $P_1$), your limit order executes and you receive consumer surplus (the lower the price, the higher the consumer surplus that you get). And now for the conclusion: When the trading facility is a call auction, the price on the limit order that you submit to the call should be your *reservation price* for the size of the order that you have chosen to submit.

When submitting his or her reservation price to a call auction, the trader does not know what the clearing price will be and thus does not know what consumer surplus he or she will realize. Assuming that the distribution of possible clearing prices is known, the expected consumer surplus is the probability weighted average of the possible consumer surplus values. In light of this, it is clear that submitting a limit order at a price below the trader's reservation price can only lower the expected value of consumer surplus because it would result in zero values replacing positive values for realizations of $P^C$ that are below $P^R$ but above the price of the improperly priced limit order. Submitting a limit order at a price higher than $P^R$ would also lower the expected value of consumer surplus because the surplus for prices greater than $P^R$ is negative. Again, we conclude that the optimal price to place on a limit order submitted to a call auction is the reservation price that is associated with the number of shares being sought for purchase.

One further question might strike you: The maximization of the expected value of consumer surplus involves maximizing the expected value of a dollar amount, not the utility value of the dollar amount, as we have done in our previous discussion of the goal of a risk averter (namely, maximize the *expected utility* of wealth or the *expected utility* of income). Have we ignored the investor's attitude toward risk? Answer: We have not. The investor's attitude toward risk, as we have already discussed, is reflected in the negative slope of his or her demand curve to hold shares, and is thus reflected in our measure of consumer surplus. The steepness of this slope reflects the decision maker's distaste for risk. There is no need to adjust further for risk.

## THE BIG PICTURE

In this chapter we have ventured away from the elegance and efficiency of a frictionless market. In doing so, we have used analytical tools set forth in Chapters 2 and 3 that we have established with reference to the frictionless market. We have seen that the introduction of information processing, trading costs, and other impediments has a profound impact on market outcomes. In the nonfrictionless world, not all mutually beneficial trades are realized, market clearing prices can deviate from equilibrium values, market participants use trading strategies when they come to a market to trade, and the very strategies that they employ can further impact market outcomes.

We considered the formulation of a trading strategy in the previous section of this chapter, but we did so in a simplified setting. First, we considered how limit orders would be placed in a highly stylized continuous limit order book market when the continuous market was the only available trading venue. Then we turned to the periodic call auction and assumed that it was the only available venue. In reality, hybrid combinations exist and more complex trading decision rules can be formulated for a call that is followed by continuous trading, and for continuous trading that closes with a call. More complex orders can be introduced. Intermediaries can be introduced. And so on and so forth. Our purpose is not to cover all of the possibilities (an impossible task), but to illustrate some of the more complex decisions that participants in real-world markets make.

When analyzing a nonfrictionless market, it is useful to operate in the context of a specific market, and the market that we have been working with is the secondary market for trading already-issued equity shares. This is a rich market to treat. It is important, it is complex, it is widely followed, and it is intriguing. One striking aspect of the equity market is the speed with which the time clock ticks, especially in an electronic environment where order submission, order handling, price determination, trade execution and reporting can take place in subsecond (e.g., millisecond) intervals. This speed, combined with the fact that prices are being set while trading is taking place, accentuates the complexities of a market process.

Further thought should be given on all of our parts (students, researchers, practitioners, and regulators) to the dynamic nature of market processes in other environments where the clock ticks at a far slower pace, where the long run is measured in terms of a year or more rather than a few hours or less, and where prices are less subject to continual change. Transaction costs are endemic to the micro markets. Short-run demand and supply curves in all markets are less elastic than their long-run counterparts because adjustments are not easy to make. Multiyear contracts, other lengthy time commitments, and protracted building and construction periods, all impact the workings of the micro markets. We return to this thought in Chapter 7.

Despite friction in the marketplace, we have suggested that Adam Smith's invisible hand is nevertheless felt in the micro markets and that the efficiency of the outcomes that the invisible hand might lead us to deserves to be treated with utmost respect. Nevertheless, problems can exist, especially in the short run after a demand or supply shock has perturbed an equilibrium solution. This reality can give rise to complex public policy issues. Understanding the operations of a micro market and appropriately harnessing its power are not simple tasks.

## CHAPTER SUMMARY

Our goal in this chapter has been to build a bridge between the theoretical properties of the frictionless environment that we consider in Chapters 2 and 3, and the realities of actual markets that we have all experienced. In doing this, we have devoted major attention to the secondary markets for already-issued equity shares. Here are the highlights:

- Friction in the economic environment is clearly reflected in one distinction that microeconomists have long made—the difference between short-run and long-run equilibrium solutions. Because of adjustment costs, short-run responses to change in economic variables are partial, and thus are less elastic than long-run responses. In considering this, we discussed adjustment costs in general and trading costs in specific. In order to hone in on trading costs, we identified eight key terms: quotations, bids, offers, limit orders, market orders, spreads, short selling, and arbitrage trading.
- Liquidity, a term that is difficult to define but quite easy to recognize, is of critical importance to a market. Illiquidity may be thought of as the flip side of the same coin that trading costs are on. Much economic analysis abstracts from illiquidity issues by assuming a frictionless environment, and that is where we started our discussion of microeconomics. But as we moved away from the perfectly liquid market, we introduced price discovery and quantity discovery as two economic objectives of major importance, and we underscored the important role that market structure plays in achieving acceptably accurate price discovery and acceptably complete quantity discovery.

  With the importance of market structure established, we turned to alternatives for market structure itself and, once again, focused on the equity markets. We saw how prices are established and trades are made in quote-driven (dealer) markets, in continuous order-driven (limit order book) markets, in periodic call auctions, and in negotiated (block) markets.
- With an overview of market structure established, our attention returned to an individual trader's participation in the market. The first item of business was to quantify a participant's gains from trading. Our measure of this, for a buyer, is "consumer surplus," a term that we

defined with the use of two demand curves (the ordinary demand curve that we have thus far been working with) and a reservation demand curve. For a seller, the analogous term is *producer surplus*. When referring to both buyers and sellers, we simply say "surplus."

For a buyer, a reservation price is the maximum that he or she would be willing to pay for a specific quantity of the good in question when the alternative is to get nothing at all. Consumer surplus, our monetary measure of the gains to trade for a buyer, is the difference between this maximum amount that he or she would pay (the reservation price) and the amount that he or she actually has to pay, times quantity. In making a trading decision, the order placer's objective is to maximize the expected value of the surplus (this is consistent with the maximization of his or her expected utility).

With the trader's objective established, we next considered his or her strategic order placement. The strategic decision was viewed as a two-step process: (1) fix the number of shares that you wish to trade, and (2) price your order. We simplified the analysis by assuming that the share size of the order has already been determined, and then focused only on price. Of particular importance is our understanding of how micro markets and their participants operate. With this in mind, we contrasted two alternative market structures, (1) the continuous limit order book market and (2) the periodic call auction. We showed that the optimal order pricing decision of our representative participant differed between the two.

- We concluded the chapter with "the big picture." Two statements merit repetition:
  1. "In the nonfrictionless world, not all mutually beneficial trades are realized, market clearing prices can deviate from equilibrium values, market participants use trading strategies when they come to a market to trade, and the very strategies that they employ can further affect market outcomes." Quite clearly, models of frictionless markets require modification and further elaboration when applied to a market where trading is costly.
  2. "Transaction costs are endemic to the micro markets. Short-run demand and supply curves in all markets are less elastic than their long-run counterparts because adjustments are not easy to make. Multiyear contracts, other lengthy time commitments, and protracted building and construction periods, all impact the workings of the micro markets." Consequently, issues such as price and quantity discovery that we have considered with respect to the equity markets may also be germane to other markets where the long run is measured not in hours or days, but in years.

## LET'S CONSIDER AN APPLICATION: WHY BID-ASK SPREADS EXIST IN SECURITIES MARKETS

In the earlier part of this chapter, we identified the bid-ask spread as a cost of trading. Later in the chapter we focused on the trading strategy an investor might employ when going to an equity market to buy or to sell shares. Can these two topics be pulled together? Could it be that investor trading strategies explain the existence of bid-ask spreads in a securities market? This is the issue that we address with our application for this chapter.*

A single list price is commonly specified for many of the goods and services that we buy. But in some cases there are two prices: a higher offer price that is extended to buyers and a lower bid price that is extended to sellers. Spreads are typical of quote-driven, dealer markets (for instance, currency dealers, securities dealers, and used car dealers) where there are public customers on both sides of the market, with some people looking to buy while others are looking to sell. In these markets, the existence of spreads is easy to understand. As an intermediary who stands between the natural sellers and the natural buyers, a dealer firm's purchase price must be lower than its selling price because the difference is the source of the dealer firm's profits.

What if buyers and sellers can meet and trade without the services of a dealer? If there is no dealer intermediary, would the spread disappear? If we turn to an order-driven market, will there still be a bid-ask spread? The answer is not obvious. Consider a limit order book such as the one shown in Exhibit 4.4. Let a very large number of participants arrive at the market, look at the book, and enter either a limit order or a market order. As more limit orders arrive, the book fills.

In Exhibit 4.4, there are no buy limits at 10.40, 10.20, and 9.90, and there are no sell limits at 10.90, and 11.30; if a fistful of new limit orders were to arrive, these "air pockets" in the book would fill in, and they would do so simply by chance. Another vacant price point in the exhibit is at 10.70. This one is within the spread. Will this price point disappear like the air pockets, or is its existence systematic? With a large enough number of participants entering limit orders, any price point between the bid and the offer would fill unless the absence of orders within the spread is systematic. That is what we have to figure out.

*Material in this application draws from Kalman Cohen, Steven Maier, Robert Schwartz, and David Whitcomb, "Transaction Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread," *Journal of Political Economy* (April 1981), 287–305.

An analytic context is needed to answer this one. Assume that a single stock is being traded in an order-driven equity market; that there is an arbitrarily large number of participants; that each participant is looking to buy or to sell just one round lot that is 100 shares (with an order size that is the same for all, we can ignore the possibility of partial executions); that price is a continuous variable (unlike the discrete prices shown in Exhibit 4.4); that participants each have a reservation price to buy or to sell shares; and that each participant has a single objective: maximize his or her expected gains from trading. With this by way of background, we ask you to address the following.

1. Consider an individual buyer with a reservation price of 12.00 who is coming to the market at a time when the best (lowest) offer on the market is 10.80. What is the buyer's gain from trade if he or she submits a market order to buy one round lot? (Remember, a round lot is 100 shares.)
2. What is the buyer's *expected* gains from trade if, instead of submitting a market order, he or she submits a limit order at 10.70 and the probability that the limit order at this price will execute is 0.95?
3. Given your answers to 1 and 2 above, which order should the buyer place, the market order or the limit order at 10.70?
4. How would your answer to 3 above change if the buyer were to place the limit order at 10.78? (Remember, for our analysis price is a continuous variable so 10.78 is a possible price point.) Let the probability be 0.96 that the limit order at 10.78 will execute.
5. Back away from the numbers for a moment. In general, how do the potential gains from trading by limit order change as the limit order gets placed ever closer to a counterpart market ask that could be hit with certainty by a market order? As the price of the limit order gets infinitesimally close to the counterpart market ask, do you agree that, if the order executes, the gain from trading by the limit order becomes infinitesimally close to the certain gain from trading by market order? Of course, if the limit order does not execute, no gain from trading is realized at all.
6. How does the probability that the limit order will execute change as the limit buy order is placed ever closer to the market ask at 10.80? Clearly, the probability that the buy order will execute is

(*Continued*)

(*Continued*)

higher, the higher is its price. But the following is key: As long asthe buy order is placed below the market ask, no matter how close to the ask it might be (even infinitesimally close), the probability that the order will execute remains discretely below one.

Some math is needed to prove the point we have just made, but the thought can be understood intuitively. Suppose your limit order is placed infinitesimally below the offer and is sitting there when some bullish news comes out. The limit sell order(s) at 10.80 will most likely execute or be withdrawn as quickly as possible, a higher ask will be posted, and your limit buy order will look more unattractive to the sellers. Does this make intuitive sense to you? Can you see that, no matter how close your buy limit is to the offer, because you have not actually *lifted* the offer with a market order, you are taking a risk, you are facing a finite probability that your order will not execute? Can you appreciate that there is a qualitative difference between trading with certainty, and not submitting a market order that would consummate the deal?

7. So, what is the story for a limit order that is placed just a hair below the counterpart sell order of 10.80 that is on the book? Quite clearly, the gain from the limit order is virtually identical to the gain from the market order, but the probability of the limit order executing does not get infinitesimally close to 1. So, here is the question: Will the buyer ever post a limit buy order infinitesimally close to an already posted offer? Reciprocally, will a seller ever post a limit sell order infinitesimally close to an already posted bid? What do you think?
8. Do you see why we have assumed that price is a continuous variable? Do you understand why the spread will never be infinitesimal when price variations can themselves be infinitesimal? Can you generalize the result and appreciate that, when price is a discrete variable, empty price points can endure within the spread (such as the one at 10.70 in Exhibit 4.4) even when an arbitrarily large number of participants are placing limit orders on the book? So, there we have it. The empty price points within the spread are systematic. Do you see that this result is attributable to the trading strategies of participants, to their appreciation of the value of trading with certainty?

9. Apply the thinking developed thus far to a bargaining situation. Suppose two people are haggling over the sale of a secondhand car that is priced in the \$8,000 to \$9,000 range. How close can their bid and offer get before the two negotiators will simply say, let's split the difference and consummate the trade? Would they keep bargaining when they are within \$200 of each other? When they are within \$100? Within \$50, or something less? Recognizing that they would both get surplus from the trade, do you agree that they will settle on a price long before the spread between them gets narrowed to \$1?
10. As you know, two quotes establish the spread: the lowest of all the ask quotes, and the highest of all the bid quotes. For the spread to endure, the best bid and the best offer must be far enough away from each other so that neither the buyer nor the seller is attracted to switch to a market order strategy, accept the counterpart quote, and trade. In other words, all posted limit orders must lie outside the "gravitational pull" of the counterpart quotes. Alternatively stated, no limit order will be posted within the "gravitational pull" of a counterpart quote. What exerts the gravitational pull? The attractiveness of trading with certainty. Can you understand why we have explained the existence of bid-ask spreads with the use of the term "gravitational pull"? Does this use of an astrological term make sense to you? Can you accept that the configuration of counterpart quotes on a limit order book must, in equilibrium, maintain sufficient space between the buys and the sells, just as the configuration of satellites, planets, and stars must preserve sufficient space between these heavenly bodies, for if they were to enter each other's gravitational field they would be drawn together?
11. Pull it all together and answer the question that we started with: Why do spreads exist in securities markets? And go further: Explain which of the following are expected to have larger spreads: A \$100 stock or a \$20 stock? A large cap, frequently traded stock, or a low cap, infrequently traded stock? A stock that experiences frequent information change, or one that does not?
12. Relax the assumption that all orders are for one round lot and consider the possibility of a large buy order and a large sell order being on the book at the same time. How would the size of the spread between the two large orders compare, on expectation, with the size of the spread between the smaller buy and sell orders?

## TraderEx SIMULATION: WHAT ARE YOUR ATTITUDES TOWARD TRADING COSTS?

Again we turn to the TraderEx simulation software that is available in *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course* to give you a further opportunity to work with some of the concepts we have developed in this text. The simulation exercise for this chapter uses the basic framework that we employ in Chapter 2. While the exercise at the end of Chapter 2 is designed to make the link between a participant's utility of income (or wealth) and his or her risk aversion more concrete, the current exercise should sharpen your understanding of the impact that trading costs have on longer-run returns. As in Chapter 2:

- You start each simulation run with $10,000 of cash in your simulated account and zero shares of a risky asset that you will buy with the $10,000.
- The simulation will run for two trading days that are separated by one year. On "day one," which is at the beginning of the year, you buy to establish your position. On "day two," which is at the end of the year, you sell to liquidate your position and turn your shares back into cash.
- Your ending cash position divided by your opening cash position, minus one, is the percentage return that you will have realized for the simulation run. Once the simulation parameters have been set, the stock's one-year return from the end of day one to the start of day two is out of your control. The computer determines this return by a random draw from a lognormal distribution that has a mean (the expected return) and variance (the risk).

For the exercise in Chapter 2, you are given a five-minute period (clock time) on day one to establish your equity position, and a five-minute period on day two to liquidate your position and see how well you have done. In Chapter 2, on both days, the limit order book is relatively deep (that is, the stock was relatively liquid). These two conditions are changed for the Chapter 4 exercise. The day one and day two trading periods are considerably longer (approximately twenty minutes of clock time), and the book is considerably thinner. With these two changes, we can redirect your attention from utility curves and risk, to trading costs and returns.

For each simulation run, set both the expected return and the standard deviation of returns equal to 10 percent. At the start of each simulation run you can set two parameters: the average time between order arrivals measured in minutes [E(ORDERS)], and the average percentage of orders that are limit orders [E(LIMITS)]. Run the simulation for each of the following two market settings:

| **Market** | **E(ORDERS)** | **E(LIMITS)** |
|---|---|---|
| A | 3.5 minutes | 80 percent |
| B | 6.5 minutes | 40 percent |

For each market, run the simulation enough times to get a sense of the distribution of returns that you might realize from that market. Then respond to the following:

1. Which of the two markets appears to be more liquid, A or B?
2. Did you get some high returns in either of these two markets (e.g., greater than 10 percent)? If so, was this because the underlying, one-year return for the stock was high, or because you traded well on day one and/or on day two?
3. Did you get some low returns in either of these two markets (e.g., less than 10 percent)? If so, was this because the underlying, one-year return for the stock was poor, or because you traded poorly on day one and/or on day two?
4. For each of the two markets, what was your average return across the set of simulations that you ran? How does your average return for each market compare with the 10 percent expected return that we set as a parameter for the simulation runs? What might explain any difference?
5. Which was the easier trading environment for you to work your orders in, market A or market B?
6. What thought did you give to each of the following while working your orders in these markets?
   - the size of the bid-ask spread.
   - market impact.
   - intraday price volatility.
   - pressure to invest your $10,000 fully on day 1, and then to sell all of your shares on day 2.

(*Continued*)

(*Continued*)

7. Do you now understand better the impact that trading costs can have on realized returns and, in the presence of these costs, that trading requires making strategic decisions?
8. With regard to the application for this chapter, did you sense in the simulation exercise that the existence of a bid-ask spread in an order-driven market can be attributed to the strategic trading decisions of participants?

## CHAPTER 5

# Supply and the Costs of Production

Thus far in this book, we have analyzed the consumer choice model, the individual consumer demand curve, and the market demand curve that the consumer choice model implies, and have interfaced the market demand curve with a market supply curve. In this chapter we focus on the supply propensities of individual firms and markets to understand their properties and the cost considerations that lie behind them. We first establish the basic analytical framework with reference to the frictionless environment and the undefined good, *X*. Then we proceed to a more concrete discussion of one specific, nonfrictionless micro market—a dealer market where equity shares are traded.

This is our game plan with regard to the generic good, *X*. Our analysis starts with the firm's production function (which describes how *X* can be produced) and the isocost lines facing the firm (which describe the alternative costs of obtaining the factors of production used in the production of *X*). Working with two factors of production (labor and physical capital) and using the production function and isocost lines, we next obtain the firm's cost curves: its total cost curve, its average cost curve, and its marginal cost curve. We distinguish between the long-run cost curves (which apply when both factors of production are fully variable), and the short-run curves (which apply when only one factor, which we take to be labor, is variable, and the other, capital, is fixed). With reference to the long-run curves, we distinguish between economies of scale (which translate into average costs decreasing with output), constant returns to scale (which translate into average costs that are constant as output increases), and diseconomies of scale (which translate into average costs increasing with output).

With the cost curves in hand, we identify the best profit output decision for the firm, and from there obtain the supply curve for a competitive firm (which is its marginal cost curve, starting at the minimum point on its average cost curve and moving up and to the right). The individual firm supply curves, when aggregated across the competitive firms, give us the supply curve for the industry. We show how the industry supply curve can reflect the existence of economies of scale that are external to the individual firms but that are realized by the industry in aggregate.

This is our game plan with regard to the equity dealer market. After providing basic information about the market (including details concerning market maker services, revenues, and costs), we present two dealer pricing models and a dynamic inventory control model, which is followed by stylized descriptions of the competitive structure of a dealer market and of a dealer firm's output and cost control. This analysis positions us to model the average cost curve for the competitive dealer firm. You will see that the average cost curve for the competitive dealer firm has the same basic shape as the average cost curve for our generic good, *X*.

## THE PRODUCTION FUNCTION FOR *X*

A firm typically produces multiple goods, and resources that are used in the production of these goods are generally supplied by other firms and by households (and also by government). Households, for instance, are suppliers of labor and financial capital. To keep it simple, in considering the supply of *X*, we take *X* to be the output of a firm that produces only one product: *X*. For the most part, we assume that a large number of firms are in the industry, that no one firm has dominant market power, and that all firms are price takers. In other words, in much of this chapter, we focus on a competitive industry.

Just as the market demand curve for *X* is the aggregate of the individual demand curves, so too is the market supply curve for *X* the aggregate of the individual firms' supply curves. For supply, as we saw for demand, aggregation is a horizontal summation process: at each price, we add the quantities brought to market at that price by each of the individual firms in the industry. The supply curve for each competitive firm is upward sloping and, accordingly, the supply curve for the industry is upward sloping (although there is an exception, which we explain later in this chapter).

Where do the individual firm supply curves come from? What do they reflect? What gives them their shapes and their locations? To answer these questions, we go behind a firm's supply curve to its cost curves. But the cost curves are not the starting point, we have to drill down further. We begin with a firm's production function and isocost lines.

## The Production Function, Isocost Lines, and the Expansion Path

A firm's production function is a technological description of how the resources that it uses, that is, its inputs into production (its *factors of production*), can be combined to produce $X$. A firm's production function underlies its production of $X$ just as a utility function underlies a household's "production" of utility.

In presenting a household's utility function, we dealt with just two consumption goods that, as generics, we called $X$ and $Y$. The advantage of dealing with just two goods is that, in doing so, we can use graphical analysis. For the same reason, in presenting the firm's production function, we consider just two factors of production: labor and physical capital (as distinct from financial capital). But at this point do not think of the two factors too concretely. Labor simply stands for some standardized human input, and capital stands for some standardized physical input. The price of labor can be visualized as a wage that is expressed as a rate per unit of time. For our analysis, the price of capital is similarly dimensioned—you could think of the price of capital as a lease rate per unit of capital per unit of time.

The production function for $X$ can be written as

$$X = f(labor,\ capital) \tag{5.1}$$

Notice that in writing this production function we are ignoring an absolutely essential input employed by any real-world firm: management skill. We can do so because this is theory, and it pays to keep things simple.

We have used the household's utility function to obtain a family of indifference curves that we interfaced with the decision maker's budget constraint to determine the optimal, utility-maximizing basket of $X$ and $Y$, given $P_X$, given $P_Y$, and given the household's income. In similar fashion, we now use the firm's production function to obtain a family of *isoquants* that can be related to the prices of the two factors of production. The isoquants and factor prices are used to determine how much labor and how much capital to employ in the production of $X$. Each individual isoquant, as its name implies ("iso" means equal and "quant" is short for "quantity"), is the locus of different labor, capital combinations that will produce an equal amount of output.

Just as the curvature of an indifference curve in utility theory reflects the complementarity/substitutability between $X$ and $Y$ in consumption, so too does the curvature of an isoquant reflect the complementarity/substitutability between labor and capital in production:

- For the firm, the relationship between the inputs of labor and capital and the output of *X* is determined by technology. For the household, the relationship between utility and the inputs of *X* and *Y* is determined by *taste*. If you think about it, you might recognize that taste incorporates a sizable technology component.
- The closer labor and capital are as complements, the more curved are the isoquants. If labor and capital are perfect complements in the production of *X*, the isoquants are strictly L-shaped and optimal factor combinations are determined entirely by technology. But if the isoquants are not strictly L-shaped, then input combinations also depend on relative input prices, just as they do in the consumer choice model.
- The closer labor and capital are as substitutes, the more linear are the isoquants. The greater the opportunities for substitution, the more important relative prices are to the efficient workings of a micro market (not only the micro market for *X*, but also the micro markets for labor and capital). We have noted this before with regard to consumer theory.

For a firm, the counterpart to a household's budget constraint is an *isocost line*. Again, "iso" means equal, and an isocost line is the locus of different labor, capital combinations that result in the same total expenditure for the firm. Individual firms in a competitive industry are also price takers in their factor markets, which means that the prices of labor and of capital are set exogenously, and each isocost line is a downward sloping, linear function, as is a household's budget constraint.

There is an important difference, however, between a budget constraint and the isocost lines. For a *household*, there is just one budget constraint; its slope is set by relative prices and its location reflects the household's income. For a *firm*, there is a *family of isocost lines*, with the slope of each line reflecting the relative prices of labor and capital. As we have said, total expenditure (cost) is constant along an isocost line, but total expenditure itself is not exogenously determined. If a firm can make a larger profit by spending more, producing more, and selling more, then it is justified in doing so. Shortly, we will see how this works.

Another similarity between isoquants and indifference curves pertains to the slopes of the two curves. With capital on the vertical axis and labor on the horizontal, the *slope* of an isoquant is (the negative of) the ratio of the marginal productivity of labor to the marginal productivity of capital, just as the slope of an indifference curve is (the negative of) the ratio of the marginal utility of *X* to the marginal utility of *Y*. The ratio of marginal productivities is called *the marginal rate of technical substitution* of labor for capital, or $MRTS_{labor \cdot capital}$ (just as the slope of an indifference curve is called *the marginal rate of substitution* of $X$ for $Y$, or $MRS_{X \cdot Y}$). We have seen that $MRS_{X \cdot Y}$ decreases as $X$ is increasingly substituted for $Y$. Similarly, $MRTS_{labor \cdot capital}$ also decreases as labor is increasingly substituted for capital.

There is a parallel argument between isocost lines and budget constraints, and it too pertains to the slopes of the curves. For both, slope is determined by (the negative of) relative prices: $P_X/P_Y$ for a budget constraint, and $P_{labor}/P_{capital}$ for an isocost line.

The goal of a household is to maximize utility, and the objective is attained by reaching the highest indifference curve possible given the consumer's budget constraint. The highest indifference curve that can be reached is the one that is tangent to the budget constraint. At the point of tangency, the slopes of the two curves are equal and hence $P_X/P_Y = MRS_{X \cdot Y}$.

Similar thinking applies to the production decision, but the wording is a bit different because total expenditures are not a constraint for the firm, as we have noted. The goal of the firm is to maximize profit, and a necessary condition for attaining this goal is to minimize cost for any level of output or, equivalently, to maximize output for any level of cost. This condition is satisfied by selecting an input combination where an isocost line is tangent to an isoquant. Vis-à-vis the point of tangency, all other points on the isoquant are being intersected by higher isocost lines, and all other points on the isocost line are being intersected by lower isoquants. Conclusion: Combining labor and capital so that $P_{labor}/P_{capital} = MRTS_{labor \cdot capital}$ is a necessary (but not sufficient) condition for profit maximization.

A mapping of the isoquant and isocost lines that we have been discussing is presented in Exhibit 5.1. Units of labor are on the horizontal axis, and units of capital are on the vertical axis. The curves labeled $X_1$, $X_2$, and $X_3$ are three representative isoquants, with $X_1 < X_2 < X_3$. The lines labeled $TC_1$, $TC_2$, and $TC_3$ are three representative isocost lines, with $TC_1 < TC_2 < TC_3$. The points labeled $A$, $B$, and $C$ are three points of tangency that show the combinations of labor and capital that are optimal in that, at each point of tangency, cost is minimized for the level of output, and output is maximized for the level of cost.

To totally convince you, go to another point such as $F$ where the $X_2$ isoquant intersects the $TC_3$ isocost line. At point $F$, an output of $X_2$ is being realized at a total cost of $TC_3$, but for an expenditure of $TC_3$, the higher $X_3$ isoquant could be reached. Or, look at it this way: at point F where output is $X_2$, a total cost of $TC_3$ is being incurred, but $X_2$ could be produced at the lower cost, $TC_2$, by operating at the point labeled B.

In Exhibit 5.1, a line labeled "expansion path" is passing through the tangency points $A$, $B$, and $C$. The *expansion path*, as its name implies, could be thought of as the path that the firm would "expand along" as it gets

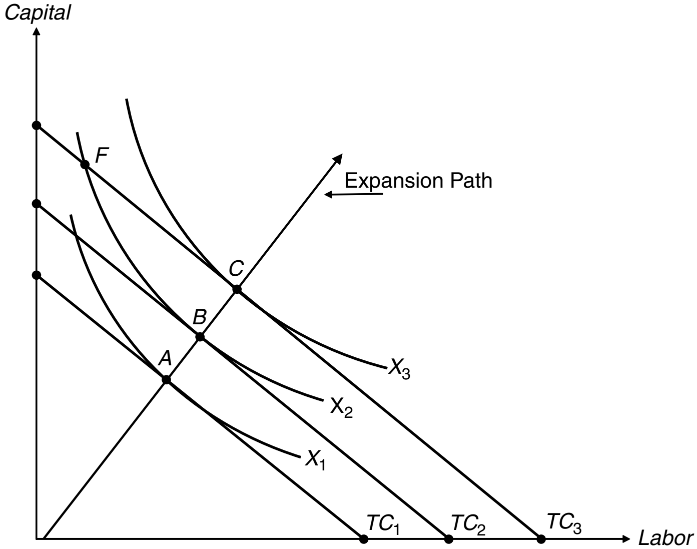


**EXHIBIT 5.1** Isoquant and Isocost Mappings for a Firm, and the Firm's Expansion Path

larger. But because our analysis is comparative static, not dynamic, we do not wish to suggest that the expansion path actually describes a growth path for the firm. Rather, it is a description of what the total cost, total output relationship is for the firm. This relationship tells us what the total production costs would be for any level of output that the firm might select.

The (positive) slope of the expansion path depends on two things:

1. **Technology:** The marginal rate of technical substitution, $MRTS_{labor \cdot capital}$, is determined by technology. If the marginal productivity of labor rises relative to that of capital, the isoquants get steeper, the points of tangency shift down and to the right, and the expansion path becomes less steep (and vice versa if the marginal productivity of capital rises relative to that of labor).
2. **The price of labor relative to capital:** If the price of labor falls relative to the price of capital, the isocost lines get flatter, the points of tangency shift down and to the right and, again, the expansion path becomes less steep (and vice versa if the price of labor rises relative to the price of capital).

Recognizing that the slope of the expansion path depends on the price of labor relative to the price of capital, we could extract information from the isoquant, isocost mappings in Exhibit 5.1 to see how the firm's demand for one of the factors of production, say labor, changes as the price of labor changes, all else being constant (where all else is the price of capital, the price of the good being produced, the price of other goods that are related in production, and technology). But we are not heading in that direction in this chapter. Our immediate concern is with obtaining the firm's supply curve of *X*.

## COST CURVES

Before we can obtain the firm's supply curve for *X*, we have to see how its costs change with its output of *X*. We can get this cost, output relationship from the information contained in the isoquant, isocost mappings. First we obtain the long-run cost, output relationship, and then we proceed to the short-run cost, output relationship.

### The Long-Run Total Cost Curve for the Firm

All that needs be done to obtain the long-run cost and output relationship for the firm is to take the labels from the isoquants and from the tangent isocost lines, and match the two labels together. We show how this works in Exhibit 5.2. In Exhibit 5.2(a), we have the expansion path from Exhibit 5.1. The location of the expansion path itself identifies the tangency points and so, rather than drawing in many isoquants and isocost lines, we have simply put hash marks on the expansion path to mark equal increments of output. If the hash marks start to bunch closer together as we move out along the expansion path, fewer and fewer additional inputs are required to realize a given increase in *X*; if the hash marks start to spread further and further apart, increasing amounts of labor and capital are required to attain equal output increments of *X*. Think of it this way: The spacing between the hash markets is like the spacing between contour lines on a topographical map that describes the shape of a mountain. The steeper is the mountain, the closer the contour lines bunch together on two-dimensional paper; the flatter is the slope, the more they spread out.

Notice that as we move out along the expansion path in Exhibit 5.2(a), the hash marks initially get closer together; the region where this is happening is identified by the bracket labeled *A*. While total costs are rising as the firm moves to higher output levels, the decreasing spaces between the hash marks in region *A* indicate that total costs are increasing at a decreasing rate. Next, when the firm is in the region identified by the bracket labeled *B*, the hash marks are equidistant along the expansion path. Here total costs

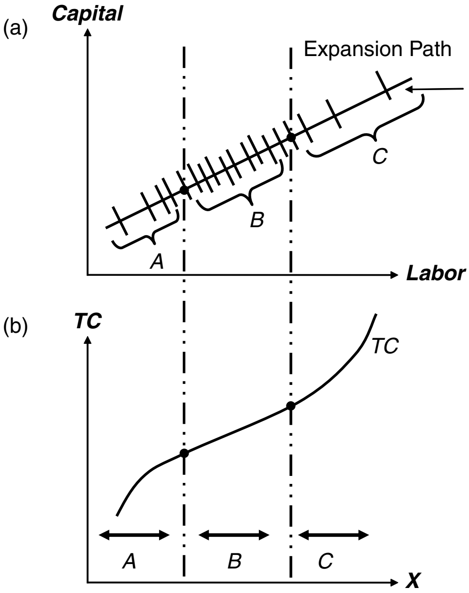


**EXHIBIT 5.2** Increasing, Constant, and Decreasing Returns to Scale Illustrated with Reference to a Firm's Expansion Path [5.2(a)] and Total Cost Curve [5.2(b)]

are rising at a constant rate. Then, further out along the expansion path, the hash marks start to widen out; the region where this is happening is labeled *C*. In this third region, costs are increasing at an increasing rate.

We have identified three cost conditions that characterize a move to higher output levels: total costs increasing at (*A*) a decreasing rate, (*B*) a constant rate, and (*C*) an increasing rate. We have another way of describing these cost conditions:

| | |
|---|---|
| A. Increasing returns to scale | (region *A*) |
| B. Constant returns to scale | (region *B*) |
| C. Decreasing returns to scale | (region *C*) |

Increasing returns to scale are expected at low levels of output. When relatively little *X* is being produced, a larger scale of operations allows the factors of production to be better coordinated and more efficiently utilized. After the economies of scale have been exploited, operations may be further scalable, but additional efficiencies are not realized and the returns to scale are constant. At some point, however, if the firm gets large enough, diseconomies of scale set in as coordinating the factors of production becomes more difficult. In fact, this must be the case for all firms in a competitive industry—the diseconomies of scale that set in as a firm gets larger limit the natural size of the individual firms, thereby enabling multiple firms to survive and preventing any one firm from gaining prominent power in the industry. In so doing, diseconomies of scale keep an industry competitive.

Let's recap. Each hash mark in Exhibit 5.2(a) should be recognized as having two labels associated with it: One shows the level of output ($X$) achieved, and the other shows the total cost ($TC$) incurred. As we move from hash mark to hash mark, $X$ increases by equal amounts (i.e., $X_2 - X_1 = X_3 - X_2 = \ldots$ ), but total cost need not. We have:

- $TC_2 - TC_1 > TC_3 - TC_2$ if the firm is realizing economies of scale.
- $TC_2 - TC_1 = TC_3 - TC_2$ if the firm is operating under conditions of constant returns to scale.
- $TC_2 - TC_1 < TC_3 - TC_2$ if the firm is experiencing diseconomies of scale.

In Exhibit 5.2(b), we graph the information concerning the total cost, output relationship contained in Exhibit 5.2(a). Total cost (the isocost line label) is on the vertical axis, and the output of $X$ (the isoquant label) is on the horizontal axis. Notice the following about the total cost curve. First, it is increasing throughout. This must be true, otherwise a lower cost way of producing, say, $X_2$, would be to produce $X_3$ and throw some of $X_3$ away (which would be absurd). Second, in region $A$, the region of increasing returns to scale, total cost is increasing at a decreasing rate; in region $B$, the region of constant returns to scale, total cost is increasing at a constant rate; and in region $C$, the region of decreasing returns to scale, total cost is increasing at an increasing rate. In long-run equilibrium, a firm in a competitive industry will always be operating in one of these three regions. Which region does your intuition say it will be, $A$, $B$, or $C$? Shortly you will see. But first, let's turn to the short-run output, cost relationship.

### The Short-Run Total Cost Curve for the Firm

We discussed the long-run, short-run distinction at some length in Chapter 4. The key difference between these two time spans is that adjustments are more difficult to make in the short run (this is in recognition of the fact that real-world markets are not frictionless environments). The clearest way to analyze the difference that adjustment costs make in the total cost, output relationship for a firm is to take a clear cut but extreme case. Assume that, in the short run, one of the factors of production is completely fixed (i.e., that the cost of immediately changing the use of this factor is infinite). We

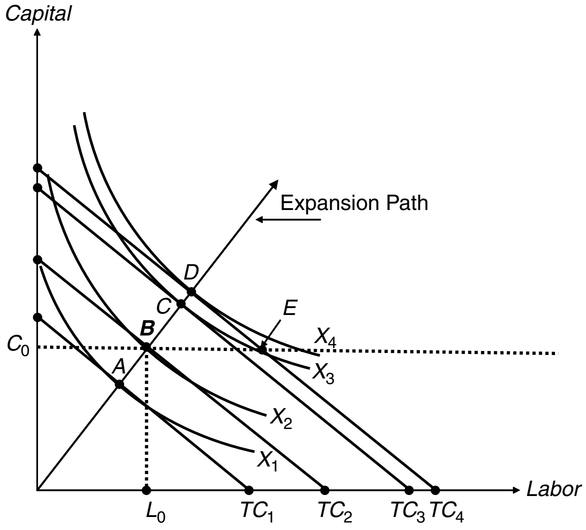


**EXHIBIT 5.3** Short-Run Output and Total Cost Relationship When One Factor of Production (Capital) Is Fixed

rather arbitrarily take capital to be the factor that is fixed in the short run. The effect of fixing capital at a given level can be seen with reference to Exhibit 5.3.

Exhibit 5.3 is the same as Exhibit 5.1 with one major exception: We have included a dotted horizontal line that starts at $C_0$ and passes through point *B*. The various points on the horizontal line show different levels of output that can be achieved by varying the labor input when the capital input is fixed at $C_0$. Perhaps the competitive firm had been operating in long-run equilibrium at point B, producing $X_2$ units of *X* (per unit of time), and then a demand shift increased the price of *X*. Can the firm respond in the short run by producing more *X*? Sure it can, and the information in Exhibit 5.3 shows how. The firm, by increasing its labor input (and, in so doing, by using its fixed capital input more intensively) could, for instance, reach the $X_3$ isoquant, as shown by point *E* in the diagram.

But point *E* is not on the expansion path. The horizontal line with the vertical intercept of $C_0$ is a constraint. To the right of point *B*, the constraint keeps the firm below the expansion path because, in this region, the expansion path can be attained only by also increasing the input of capital. To the left of point *B*, the situation is just the opposite; the constraint keeps the firm above the expansion path because, in this region, the expansion path can be attained only if the input of capital is also cut back. The constraint is not effective only at point *B*, where the horizontal line crosses the expansion path.

Another new line is also included in Exhibit 5.3: The isocost line labeled $TC_4$ that passes through point *E*. Isocost line $TC_4$ is not tangent to isoquant $X_3$ (moving from left to right it crosses it from above), and thus the short-run cost of producing $X_3$ units is higher than the long-run cost that would be achieved if capital, along with labor, was freely variable. Point *C* in the diagram is the long-run optimal solution for an output of $X_3$ (point *C* lies on the isocost line $TC_{3,}$, which is below and to the left of $TC_4$).

In the long run, $X_3$ can be produced at $TC_3$, which is less than $TC_4$, but the long-run solution is not available in the short run. The constraint (the adjustment friction) has pushed costs up. The bottom line is that a short-run total cost curve is everywhere above its long-run counterpart except at one point, the point where the short-run isocost line is the same as the long-run isocost line. At the point of tangency, the input of labor is being properly combined with the $C_0$, fixed input of capital. When the firm is at *B* in Exhibit 5.3, it is operating at such a point.

Not surprisingly, as a demand shift increases the price of *X*, with costs being higher in the short run, the firm will not expand output as much in the short run as it will in the long run. In other words, the short-run supply curve is less elastic than its long-run counterpart. But we are getting ahead of the story.

## Average Cost and Marginal Cost

The information contained in a total cost curve can be translated to show how average cost and marginal cost change as output increases. We do this with reference to Exhibits 5.4 and 5.5. For the discussion in this subsection, we need not distinguish between long-run and short-run costs because the translation is qualitatively the same for both. The only proviso to keep in mind is that fixed costs apply in the short run, but not in the long run (in the long run, all costs are variable). Accordingly, in considering Exhibits 5.4 and 5.5, recognize that we are dealing with variable costs. We do so for three cost measures: total variable costs, average variable costs, and marginal costs (which, of course, are always variable).

Exhibits 5.4(a) and 5.5(a) display the same total cost curve, *TC*. This is essentially the same curve as the one shown in Exhibit 5.2. The only difference is that, for simplicity, we have collapsed the region of constant returns to scale to one point (we lose nothing by doing this; the region is still represented in the graph even though it is only one point). Average cost and marginal cost are defined as follows:

- **Average cost:** For any output *X*, the average cost associated with the value of total cost, *TC*, is the total cost *divided by X*, or *TC/X*. Average

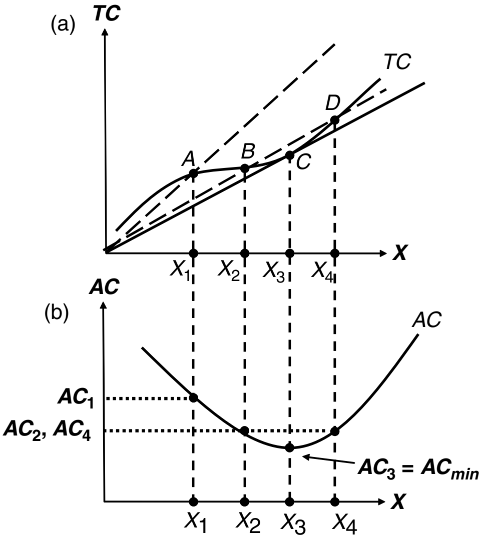


**EXHIBIT 5.4** Relationship between a Total Cost Curve [5.4(a)] and Its Associated Average Cost Curve [5.4(b)]

cost at any value of $X$ along the total cost curve is given geometrically by the slope of a ray from the origin to that point on the total cost curve.

- **Marginal cost:** For any output $X$, the marginal cost is the *change* in total cost divided by the *change* in $X$, or (assuming very small changes) $\frac{dTC}{dX}$. Marginal cost at any value of $X$ along the total cost curve is given geometrically by the slope of the line that is tangent to the total cost curve at that value of $X$.

The values of average cost and marginal cost are related to each other. We obtain this relationship by writing

$$TC = AC \times X \tag{5.2}$$

Working with very small changes, from equation (5.2) we can write the *change of TC* as the amount by that $X$ changes times a constant value of $AC$, plus the amount by which $AC$ changes times a constant value of $X$, or

$$dTC = AC \times dX + dAC \times X \tag{5.3}$$

[You might recognize that equation (5.3) is a total derivative.] Dividing equation (5.3) by $dX$ gives the marginal cost,

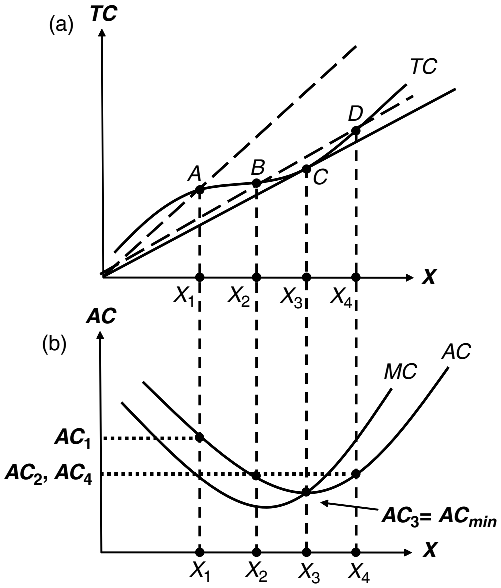


**EXHIBIT 5.5** Relationship between a Total Cost Curve [5.5(a)] and Its Associated Average and Marginal Cost Curves [5.5(b)]

$$MC = \frac{dTC}{dX} = AC + \frac{dAC}{dX}X \qquad (5.4)$$

Factoring out $AC$ on the right hand side of equation (5.4) gives

$$MC = AC(1 + \eta_{AC.X}) \qquad (5.5)$$

where $\eta_{AC.X} = \frac{dAC}{dX}\frac{X}{AC}$ is the elasticity of average cost with respect to $X$. We learn the following from equation (5.5):

- When $AC$ is falling as $X$ increases $\left(\frac{dAC}{dX} < 0\right)$, $\eta_{AC \cdot X}$ is negative and $AC$ is greater than $MC$.
- When $AC$ is rising as $X$ increases $\left(\frac{dAC}{dX} > 0\right)$, $\eta_{AC \cdot X}$ is positive and $AC$ is less than $MC$.
- When $AC$ is constant as $X$ increases $\left(\frac{dAC}{dX} = 0\right)$, $\eta_{AC \cdot X}$ is zero and $AC$ equals $MC$.

We now go to the graphs.

Exhibit 5.4 shows visually how we extract information about the average cost curve from the total cost curve. In the exhibit, $B$ marks an inflection point, the point where total cost changes from increasing at a decreasing rate to increasing at an increasing rate. The slope of a ray from the origin to any point such as *A* on the *TC* curve (the ray is the dashed line) identifies the average cost of producing the associated level of output. As the point moves from *A* to *B* and all the way to *C*, the slope of the ray decreases and therefore average cost decreases. At point *C*, the ray (the solid line) is tangent to *TC* and thus the slope of the ray and hence average cost are at a minimum. From *C* to *D*, and beyond, the slope of the ray from the origin is increasing as we move further out along the curve. We invite you to sketch in another dashed line to confirm this. Conclusion: as we have noted, at point *C* average cost is at a minimum, and beyond *C* average cost is rising.

The information in Exhibit 5.4 is repeated in Exhibit 5.5 with the marginal cost curve included in Exhibit 5.4(b). Marginal cost at any point on the *TC* curve is given by the slope of the tangent to the *TC* curve at that point. Throughout the concave portion of *TC* up to the inflection point labeled *B*, marginal cost is falling; at *B* (where $X = X_2$) marginal cost is at a minimum; and beyond *B*, we are in the convex portion of the *TC* curve and marginal cost is rising. In accordance with equation (5.5), we see that when the *AC* curve is falling, *MC* is less than *AC*; when the *AC* curve is at a minimum, *MC* equals *AC*; and when the *AC* curve is rising, *MC* is greater than *AC*. This makes intuitive sense: As long as the cost of producing the marginal unit is less than the average cost, the average cost is being pulled down, regardless of whether the *MC* curve itself is rising or falling. Then, when the cost of the marginal unit equals the cost of the average unit, average cost must be constant. Finally, when the cost of producing the marginal unit is greater than the cost of producing the average unit, the average cost must be rising.

This completes the picture concerning how average costs are related to marginal costs. Notice that both the *AC* and *MC* curves are U-shaped. We achieved the simple U-shape for both curves by collapsing the region of constant returns to scale to a single point.

Note that economies of scale are defined with regard to the behavior of average cost, not marginal cost. As long as average costs are falling, the firm is enjoying economies of scale, even though costs on the margin might be heading up. And, as long as average costs are rising, the firm is operating under conditions of decreasing returns to scale. Again we ask, under what condition does your intuition tell you that a competitive firm in long-run equilibrium will be operating: economies or diseconomies of scale?

### Average Fixed Costs

The relation between average fixed cost (*AFC*) and output is straightforward. Letting *FC* be the total amount of fixed cost, we have

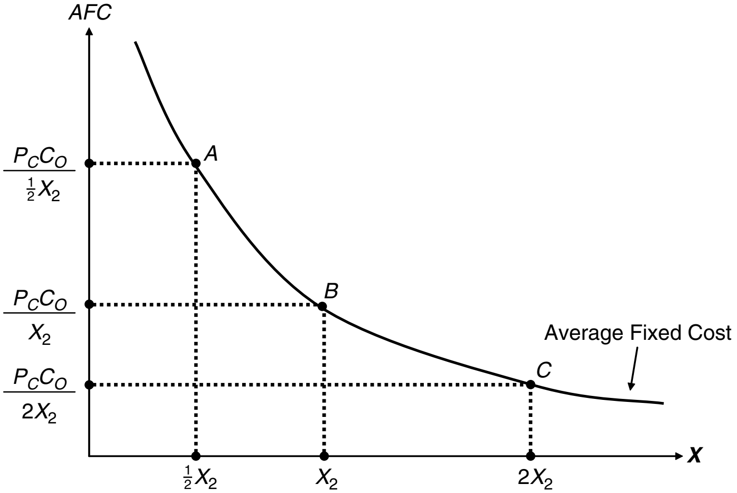


**EXHIBIT 5.6** Average Fixed Cost Curve

$$AFC = \frac{FC}{X} \tag{5.6}$$

We show the relationship given by equation (5.6) graphically in Exhibit 5.6. To tie the graph into the information contained in Exhibit 5.3, we have made point *B* in both exhibits compatible. In Exhibit 5.3, point *B*, which is on the expansion path, identifies an output of $X_2$ that is achieved with an input of $C_0$ units of capital (the fixed factor of production) and $L_0$ units of labor (the variable factor of production). Let $P_C$ be the cost per unit of capital. Then $FC = P_C C_0$ and, for the output $X_2$, $AFC = \frac{P_C C_0}{X_2}$, as shown by point *B* in Exhibit 5.6. Two other points are also shown in Exhibit 5.6. At point *A*, output is $\frac{1}{2}X_2$, and at *C*, output is $2X_2$. We see that when output is cut in half *AFC* doubles, and that when output is doubled *AFC* is cut in half. This is the simple result of *FC* being constant. This means that the area under the graph (i.e., the area of the rectangle defined by the value of *X* and the associated value of *AFC* at *X*) is the same at point *A*, as it is at point *B*, as it is at point *C*. In technical parlance, the *AFC* curve is a rectangular hyperbola.

## Average and Marginal Cost Curves in the Long Run and Short Run

In this section we deal with the shapes and relative positions of long-run and short-run average cost (*AC*) and marginal cost (*MC*) curves. The analysis is rather technical, so fasten your seat belt and let's get started.

Let's first identify the components of the average costs. Total costs (*TC*) are the sum of fixed and variable costs. Thus, given that average cost is total cost divided by output *X*, average costs can be represented as a sum of average fixed costs (*AFC*) and average variable costs (*AVC*). Using *SR* for "short run" and *LR* for "long run," we have the following: In the short run, $SRAC = SRAFC + SRAVC$, and in the long run, $LRAC = LRAVC$ (that is, $LRAFC = 0$ because there are no fixed costs in the long run).

We first consider the shapes and the relative positioning of the three average cost curves (*AC*, *AFC*, and *AVC*) in the long run and in the short run, and then turn to the shape and relative positioning of the *MC* curves.

- **Average cost curves in the long run:** The *LRAC* curve is essentially the *AC* curve in Exhibit 5.4(b), and it is U-shaped. Therefore, both *LRAC* and *LRAVC* curves are U-shaped because these two curves are, in fact, the same (as we have said, there are no fixed costs in the long run).
- **Short-run average fixed cost (*SRAFC*) curve:** Exhibit 5.6 shows the shape of the *SRAFC*; it is a negatively inclined, rectangular hyperbola. So, what do you expect the shape of *SRAC* and *SRAVC* to be? We answer this in the next two bullet points, respectively.
- **Short-run average cost (*SRAC*) curve:** Because the firm has the ability to make adjustments in the long run that are not possible in the short run, short-run costs (either total or average for a given level of output) must be equal to or higher than the comparable long-run costs.[1] Recognizing that
    - the *SRAC* curve is with one exception above the *LRAC* curve,
    - the *LRAC* curve is U-shaped, and
    - the *SRAC* and *LRAC* curves share one point in common, the point where the short-run total costs equal the long-run total costs, such as at point B in Exhibit 5.4, it follows that the *SRAC* is U-shaped, just as is the *LRAC* curve, and that it lies above *LRAC* except for the one point where the two curves are tangent.

---

[1] This intuitive statement can be supported more rigorously with reference to Exhibit 5.3. Consider point *B* in that Exhibit. At point *B*, the total cost $TC_2$ is the minimum cost of producing $X_2$ in the long run. Changing the output in the short run by changing the labor input with capital fixed at $C_0$ results in total costs that are higher in the short run (when the capital input is $C_0$) than in the long run (when the capital input can also be changed). For example, point *E* shows that producing $X_3$ with $C = C_0$, incurs a total cost of $TC_4$, while point *C*, which lies on the same isoquant as *E* (and therefore corresponds to the same level of output), shows that, when $X_3$ is produced with an optimal input of capital, the total cost is $TC_3 < TC_4$.

- **Short-run average variable cost (*SRAVC*) curve:** We have established the relationship between the three short-run average cost curves: $SRAC = SRAVC + SRAFC$. Since *SRAC* is U-shaped and *SRAFC* is a rectangular hyperbola, the *SRAVC* curve (which can be thought of as a difference between the *SRAC* and *SRAFC* curves) must also be U-shaped.

To summarize, we have thus far established the following:

- *LRAC* curve (which is the same as *LRAVC* curve) is U-shaped.
- *SRAC* curve is U-shaped and lies above the *LRAC* curve at all values of *X* except for the point where it is tangent to *LRAC*.
- *SRAC* can be represented as a sum of the *SRAFC* curve, which is a rectangular hyperbola, and *SRAVC*, which is U-shaped.

You might still ask why the *AC* curves, both long-run and short-run (*LRAC* and *SRAC*), are U-shaped. The answer is most simply seen with reference to the long-run total cost (*TC*) curve in Exhibit 5.4.

- The *TC* curve is concave (from below) in the region of increasing returns to scale, and convex (from below) in the region of decreasing returns to scale.
- The point along the curve where *TC* changes from concave to convex is the inflection point (point *B* in Exhibit 5.4). The minimum point on the *AC* cost curve (which is determined geometrically by the point on the *TC* curve where the ray from the origin is tangent to *TC*) must therefore be in the convex region of *TC* to the right of the inflection point. This point of tangency is point C in Exhibit 5.4.
- The average costs are falling in the concave region and in part of the convex region, reach a minimum at point *C*, and increase in the remainder of the convex region.

Consequently, AC falls, bottoms out, and then rises, and we call this pattern U-shaped.

- **Marginal cost curves:** What about the shape of an *MC* curve? For this one, return to Exhibit 5.5. As shown in the exhibit, marginal costs are falling throughout the region where *TC* is concave (from the origin to point *B*), are at a minimum at the inflection point (point *B*), and are rising throughout the region where *TC* is convex (to the right of point *B*). We thus see that the marginal cost curve is U-shaped, as is the *AC* curve.

Now onto the *MC* curve's location vis-à-vis the *AC* curve. From equation 5.5 and the subsequent comments, it follows that an *MC* curve is

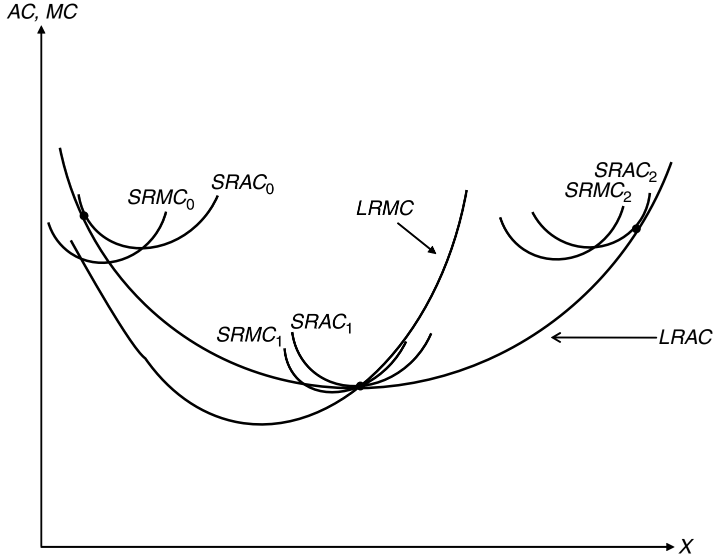


**EXHIBIT 5.7** A Long-Run Average Cost Curve (an Envelope Curve) and Marginal Cost Curve, and Their Relationships to a Family of Short-Run Average and Marginal-Cost Curves

below its associated U-shaped *AC* curve in the region where the *AC* curve is falling, is above it in the region where the *AC* curve is rising, and that it crosses the *AC* curve from below at the point where the *AC* curve is at its minimum.

Exhibit 5.7 shows the *LRAC* and the *LRMC* curves. The *LRAC* curve is identical to the *AC* curve shown in Exhibit 5.4(b). In addition, three *SRAC* curves and three *SRMC* curves are also shown in Exhibit 5.7. The *SRAC* curves are qualitatively identical to their long-run counterparts, although the short-run U-shapes are more pronounced. Of particular importance is the location of the *SRAC*s relative to the *LRAC*. Each *SRAC* is tangent to *LRAC* at one point, and is above it otherwise. This can be understood by looking again at Exhibit 5.3. Let's first focus on point *B* in Exhibit 5.3. To minimize its long-run costs, the firm operates on its expansion path. At point *B*, which is on the expansion path, the firm is combining $L_0$ units of labor with $C_0$ units of capital. In the name of "concreteness," think of $C_0$ as representing a plant of a given size (call it the "$C_0$ plant") that is being operated with $L_0$ workers.

In Exhibit 5.7, we show the *SRAC* curve that is associated with the $C_0$ plant; it is the curve labeled $SRAC_0$. Moving to the right in Exhibit 5.7, the next average cost curve shown is labeled $SRAC_1$; this curve shows how average cost changes when the firm combines its variable input, labor, with a $C_1$ plant (which is larger than the $C_0$ plant). Moving yet further along the *LRAC* curve brings us to an even larger plant, the $C_2$ plant, for which the short-run average cost curve is $SRAC_2$.

Capital is freely and continuously variable in the long-run, and we can arbitrarily pick any point along the *LRAC* curve and identify the input of capital that is associated with it. In other words, a family of *SRAC* curves exists; in Exhibit 5.7, we have selected just three curves to illustrate that family. We started our exposition with the *LRAC* and then identified the *SRAC* curves associated with it. Alternatively, we could have plotted a sizable number of *SRAC* curves and then identified the *LRAC* curve that contains them. By this second procedure we can, using mathematical parlance, identify the *LRAC* as an *envelope curve*. A mathematical procedure exists for solving for the long-run envelope function, given the family of short-run functions.[2]

We show three different tangency positions in Exhibit 5.7:

1. ***LRAC* decreasing:** In the region where the *LRAC* curve is decreasing, the point of tangency is to the left of the minimum point on the *SRAC* curve. In this region along the *LRAC* curve, the firm is enjoying increasing returns to scale; accordingly, it builds a somewhat larger plant (to better exploit the scale economies) and then underutilizes it.
2. ***LRAC* constant:** In the region where the *LRAC* curve is constant, the firm is experiencing constant returns to scale, and the point of tangency is at the minimum point on the *SRAC* curve. In this region along the *LRAC* curve the firm builds the appropriately sized plant and utilizes it perfectly.
3. ***LRAC* increasing:** In the region where the *LRAC* curve is increasing, the point of tangency is to the right of the minimum point on the *SRAC* curve. In this region along the *LRAC* curve, the firm is incurring decreasing returns to scale; accordingly, it builds a somewhat smaller plant (to temper the scale diseconomies), and then overutilizes it.

---

[2] With reference to Exhibit 5.7, note that long-run marginal cost (*LRMC*) equals short-run marginal cost for a given plant (e.g., $SRMC_0$) at the output ($X$) where the short-run average cost curve for the plant (e.g., $SRAC_0$) is tangent to the long-run average cost curve. Equality is explained by the fact that, at the point of tangency, the two average costs are the same, $X$ is the same, and thus total cost is the same, while short-run total cost is greater than long-run total costs for both smaller and larger values of $X$. Thus the short-run total cost curve is tangent to the long-run total cost curve at $X$, which means that their slopes (which reflect marginal cost) are the same.

## DETERMINING THE BEST PROFIT OUTPUT

We have analyzed the least cost combination of input factors in the production of $X$, and have translated this information into the firm's cost curves. One more step has to be taken before we can identify the supply curve of $X$ for the competitive firm: the best profit output for the firm must be determined, given the location of its revenue and cost curves.

A firm's profits are the bottom line of its income statement. You start with total revenue, subtract out all expenditures, and what remains is profits. Here is a super-simple rendition of an accountant's income statement:

$$TR - TC = \Pi$$

where $\Pi$ is the symbol that microeconomists commonly use for profits.

All three variables, $TR$, $TC$, and $\Pi$, are functions of $X$. Go to Exhibit 5.8. We show the firm's total revenue and total cost curves ($TR$ and $TC$, respectively) in Exhibit 5.8(a). In Exhibit 5.8(b) we show the associated

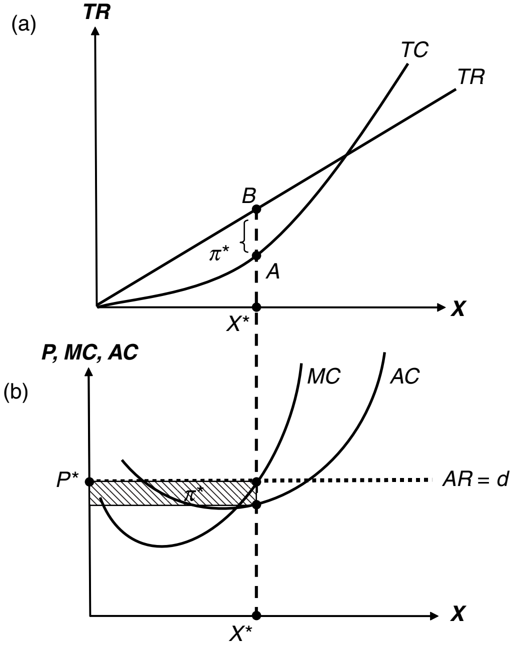


**EXHIBIT 5.8** Determination of a Firm's Best Profit Output Using Total Cost and Total Revenue Curves [5.8(a)] and Marginal Cost, Average Cost, and Average Revenue Curves [5.8(b)]

marginal cost and average cost curves (*MC* and *AC*, respectively) and the firm's average revenue function ($AR = d$). Notice that the *TR* function is linear: The reason is that the firm, being small relative to the aggregate of all firms in the competitive industry that we are modeling, is a price taker. Linearity of *TR* in Exhibit 5.8(a) translates into *AR* being constant at the level $P^*$ (the market-determined price of *X*) in Exhibit 5.8(b).

The horizontal *AR* line is the demand curve faced by the firm, which is why it is labeled $AR = d$. Being horizontal, the demand curve is infinitely elastic: at a price one tick above $P^*$ the firm would sell nothing, while at $P^*$ (or below) it could sell (cost considerations aside) an unbounded amount (until, that is, it becomes large enough to have market power). A rising marginal cost curve, however, would keep the competitive firm from ever producing an amount large enough to impact the market price (more on this in Chapter 6).

With the firm's demand curve being infinitely elastic, the *AR* curve and the *MR* curve are identical. This equality can be understood intuitively: if average revenue is constant, then the marginal revenue brought in by an additional unit of *X* must be the same for all *X*, it is the price of *X*, which is $P^*$ (and $P^*$ is average revenue). Here is another way to look at it: to sell more *X*, the competitive firm does not have to lower its price on all of its *intra-marginal units* in order to sell one more marginal unit. All it has to do is to produce the larger quantity and bring it to market.

Here is a more formal demonstration of the equality between *AR* (which we will now refer to as "*P*") and *MR*. The demonstration follows the procedure that we used to obtain the relationship between marginal cost and average cost [look at equations (5.2) – (5.5)]. First write

$$TR = P \times X \tag{5.7}$$

Working with very small changes, from equation (5.7) we can write the *change of TR* as *P* times the change in *X*, plus *X* times the change in *P*, or

$$dTR = P \times dX + X \times dP \tag{5.8}$$

Dividing equation (5.8) by *dX* gives the marginal revenue,

$$MR = \frac{dTR}{dX} = P + X\frac{dP}{dX} \tag{5.9}$$

Factoring out *P* on the right-hand side of equation (5.9) we get

$$MR = P\left(1 + \frac{1}{\eta_{X \cdot P}}\right) \tag{5.10}$$

where $\eta_{X \cdot P} = \frac{dX}{dP}\frac{P}{X} < 0$ is our familiar *elasticity of X with respect to P*. It follows from equation (5.10) that, for infinite elasticity (that is, for $\eta_{X \cdot P} = -\infty$), $MR = P$. It also follows from equation (5.10) that $MR$ is positive only if the absolute value of $\eta_{X \cdot P}$ is greater than 1.

With these revenue and cost relationships, the necessary condition that must be satisfied for the firm to maximize its profits is now within easy reach. In Exhibit 5.8(a), profit is the vertical distance between *TR* and *TC*. When *TC* is below *TR* and is rising at a slower rate than *TR*, the distance between the two curves is increasing; when *TC* is below *TR* but is rising at a faster rate than *TR*, the distance between the two curves is decreasing. Thus, profits are maximized where the two curves are increasing at the same rate. That is, profits are maximized where the line that is tangent to the *TC* curve is parallel to the *TR* curve. This condition is satisfied at the quantity $X^*$ in Exhibit (a). The profit maximizing output is $X^*$, and profit is the vertical distance $\Pi^*$ between points *A* and *B*.

What can we say about the relationship between marginal cost and marginal revenue when the tangent to the *TC* curve is parallel to the *TR* curve? Answer: the two are equal. In Exhibit 5.8(b), we see a second way to identify the profit maximizing output of the firm: it is the output level $X^*$, where $MC = MR = AR = P^*$. In Exhibit 5.8(b), profits are given by the area of the shaded rectangle defined by $(P^* - AC)X^*$. The intuitive explanation of the $MC = P^*$ condition is straightforward: if an incremental unit of *X* sells for more than it costs (on the margin) to produce, go for it; if the incremental unit costs more to produce than it sells for, forget it; if the marginal revenue just equals the marginal cost, the firm is operating on the fine edge, right at the profit maximizing level.

## SUPPLY

There you have it, both of the profit-maximizing optimality conditions for the firm have been identified: (1) employ a least cost combination of inputs (which in the long run is achieved by combining factors of production so that the ratio of their marginal products equals the ratio of their prices), and (2) operate at an output level where marginal cost equals marginal revenue. The second condition enables us to identify the firm's supply curve. We make the identification by seeing what happens to the best profit output as the demand curve faced by the competitive firm shifts. This is a good way to establish the supply curve. The curve is defined by the alternative equilibrium points that are realized as the firm's demand curve shifts, *all else constant* (or, as we like to say, *ceteris paribus*). Let's find it. To do so most easily, we focus on the long-run cost curves and, in so doing, avoid having to distinguish between fixed and variable costs.

## The Firm's Supply Curve

Exhibit 5.9 presents three alternative demand curves: $d_1$, $d_2$, and $d_3$. For each demand curve, one or more output levels ($X_1, \ldots, X_5$) are located where $MC = P$. Are all of these output levels (identified in the figure by points $A, \ldots, E$) on the firm's supply curve? At points A, B, and C, where output is, respectively, $X_1$, $X_2$, and $X_3$, price (and therefore average revenue) is below average cost, a lineup that translates into negative profits. The firm could do better. At $P_3$, *E* dominates *A*; at $P_2$, *D* dominates *B*; and at $P_1$, producing nothing dominates C. By producing nothing, the firm can realize zero profits (remember, there are no fixed costs in our long-run analysis).

This means that points *A*, *B*, and *C* are not on the firm's supply curve. At point *D*, where price is $P_2$, the demand curve is tangent to the minimum point on the firm's long-run average cost curve. Here, the firm would just break even. Profits would be zero, and the firm would be indifferent between producing $X_4$ and producing nothing. At point *E*, where price is $P_3$, the firm would equate marginal cost and marginal revenue by producing at the $X_5$ level. At this point, the firm would be maximizing its profits, given the price of *X*. Let's repeat: Given that the price of *X* is $P_3$, the firm could not operate as profitably at any other output rate (including zero) as it can at $X_5$, and so $X_5$ is the best profit output for the firm when the market price of *X* is $P_3$.

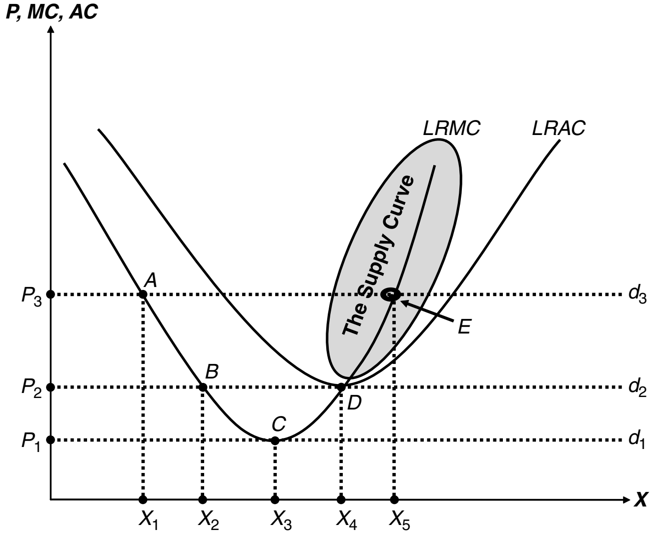


**EXHIBIT 5.9** The Long-Run Supply Curve of a Competitive Firm: The Portion of the Long-Run Marginal Cost Curve above and to the Right of the Minimum Point of the Long-Run Average Cost Curve

This is where we stand: Points *A*, *B*, and *C* are not on the firm's supply curve, while points *D* and *E* are. Point *E* was arbitrarily picked from the long-run marginal cost curve above the minimum point on the long-run average cost curve. *E* could be located anywhere along this portion of the *LRMC* curve. And so we have found the long-run supply curve of the competitive firm: It is the portion of the long-run marginal cost curve that starts at the minimum point of the long-run average cost curve and, from there, rises above it. The part of the *LRMC* that is the supply curve is highlighted by the oval that contains the words, "the supply curve."

Having identified the supply curve of the perfectly competitive firm, we can answer the question that we posed earlier in this chapter: Under what production conditions will the perfectly competitive firm operate: (a) increasing returns to scale, (b) constant returns to scale, or (c) decreasing returns to scale? We now know that the best profit output is always in the region where marginal cost is rising and is equal to, or is above, the average cost curve. In this region, the average cost curve is rising. The answer, therefore, is (c) *decreasing returns to scale*.

What about (b) *constant returns to scale*? The constant returns to scale solution is a rather special case. No profits are realized when (with reference to Exhibit 5.9) price is $P_2$ and demand is tangent to the *LRAC* curve. Aside from this point, the firm operates under conditions of decreasing returns to scale. So, the real winner is (c) *decreasing returns to scale*. To the layman, operating under decreasing returns to scale might appear to be inefficient. Why are we producing so much, one might ask, if we are getting less and less back on the margin? The layman would be wrong. Quite the opposite is true for a competitive industry: constraining output when increasing amounts of production are being realized on the margin is what would be inefficient.

Did your intuition lead you to this answer? It is an interesting one. A firm that brings outputs such as $X_1$, $X_2$, or $X_3$ to market is operating inefficiently; it is not fully exploiting the scale economies available to it. The firm should operate more intensively, it should expand into the region of diminishing returns to scale. Only then can it be maximizing profits.

Knowing the supply curve of the firm, we can now explain why the price and quantity axes have been "reversed" for graphical demand curve analysis (that is, why the dependent variable, quantity, is on the horizontal axis and the independent variable, price, is on the vertical axis). The reason is that, for cost analysis, the dependent variable is cost (and, more to the point, marginal cost), and cost is on the vertical axis. Thus price is on the vertical axis for the supply curve. We must do the same for the demand curve so that it can be related, geometrically, to the supply curve. Thus we keep price on the vertical axis for demand as well as for supply.

## Externalities and the Supply Curve

Another factor can affect the supply curve of a firm in the competitive industry, and thus the industry's aggregate supply curve. That factor is an economy of scale that is external to the firm but internal to the industry. We explain how this works with reference to a *positive* economy of scale.

A competitive industry may be able to operate at lower cost as it grows in size. When it comprises a larger number of firms, an industry may be able to acquire its factors of production at lower cost; it may realize benefits from research and development expenditures that are financed by the industry as a group; it might, having more political clout, benefit from the passage of more favorable government legislation; it might deliver more and better information to its members and in other ways improve its infrastructure; and so forth. We add a few more reasons later in this chapter when we consider the supply of market maker services in a competitive dealer market. These benefits accrue on the industry level (they are *internal to the industry*) but no one firm is able to realize them individually (they are *external to the decisions of the individual firms*). However, these economies certainly benefit the individual firms. They lower the cost curves of the individual firms and, in so doing, shift the individual firm supply curves down and to the right. We can see what is involved with reference to Exhibit 5.10.

Exhibit 5.10(a) displays two average cost curves and their associated marginal cost curves for a representative firm. The horizontal axis in the figure identifies the individual firm's output of $X$ as $X_{FIRM}$. Exhibit 5.10(b) displays three industry aggregate supply curves; the horizontal axis in the figure identifies the aggregate industry output of $X$ as $X_{INDUS}$. Because the industry enjoys economies of scale that are external to the individual firms, the locations of the average and marginal cost curves for the firm depend on the output level at which the industry is operating. Exhibit 5.10(a) displays two sets of long-run cost curves: as the labels show, one set applies when $X_{INDUS} = X_1$, and the other set applies when $X_{INDUS} = X_2$. Given that $X_1 < X_2$ and that the industry is enjoying scale economies, the cost curves labeled $X_{INDUS} = X_1$ are higher for the firm than the cost curves labeled $X_{INDUS} = X_2$.

The two steeper industry supply curves shown in Exhibit 5.10(b) have labels attached to them that are comparable to the labels used in 5.10(a). Each of these industry supply curves is the horizontal aggregate of the individual firm supply curves with the comparable labels. Each firm level supply curve is the firm's marginal cost curve starting at the minimum point on its average cost curve and rising above it as we go to higher output levels. Each marginal cost curve for a firm has a label that identifies the industry output

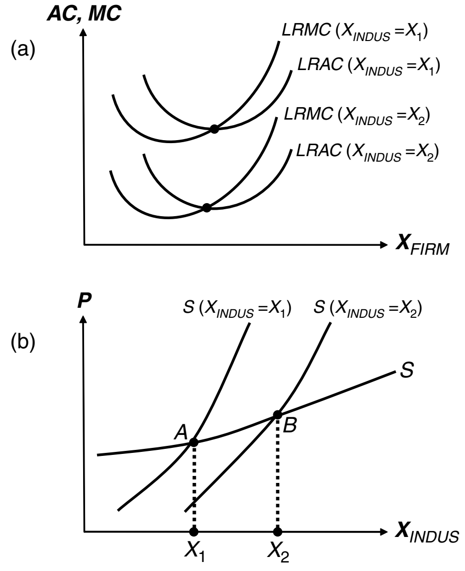


**EXHIBIT 5.10** The Long-Run Cost Curve of a Competitive Firm [5.10(a)], and the Supply Curve of a Competitive Industry [5.10(b)] under Economies of Scale That Are External to the Firms but Internal to the Industry

($X_{INDUS}$) for which the cost curve applies, and the aggregate of these curves carries the same label. Thus, in 5.10(b), the supply curve that is steeper and further to the left is labeled S($X_{INDUS} = X_1$); this is what the industry supply curve would be if the costs of production associated with an industry output of $X_1$ were to apply to all levels of industry output. Similarly, the second industry supply curve that is labeled S($X_{INDUS} = X_2$) shows what the industry supply curve would be if the costs of production associated with an industry output of $X_2$ were to apply to all levels of industry output.

The two relatively steep curves in Exhibit 5.10(b) represent a family of curves. With scale economies that are outside the reach of an individual firm and are, therefore, external to it, one aggregated curve is applicable for *each level of industry output*, but only one point on each of the aggregated curves is relevant: the one point that corresponds with the assumption upon which the entire curve is based. Point *A* in Exhibit 5.10(b) is such a point—at *A*, the marginal cost curve has the label S($X_{INDUS} = X_1$) and, at *A*, the industry's output is in fact $X_1$. Point *B* is another such point—at *B*, the marginal cost curve has the label S($X_{INDUS} = X_2$) and, at *B*, the industry's output is in fact $X_2$.

We can now identify the industry's supply curve of *X* under the condition of scale economies that are external to the individual firms. The industry supply curve is the flatter curve labeled "*S*" (no further label attached) that passes through points *A* and *B*. Notice that, being flatter, the supply curve S is more elastic than the curves that it is cutting across. This makes intuitive sense. Economies of scale that are external to the individual firms should increase the supply response of the industry in aggregate. It is in fact possible, if the scale economies are strong enough, for the *S* curve to even be downward sloping. Try reproducing the figure in Exhibit 5.10(b) and see if you can get this result (or wait until we return to this later in the chapter).

## When the *MC* Curve Is Not a Supply Curve

The answer, "It is all demand and supply," is commonly given in news broadcasts when a topic such as the rising or falling price of oil is being discussed and the anchor has asked "Why?" Although it might sound like an uninformative cop-out, the statement may well be correct. We have been discussing demand and supply, we have seen how the buying pressure of people who use *X* can interact with the selling pressure of firms that produce *X*, and how the interaction between the two establishes efficient price and quantity values for *X*. Market demand and supply curves have played a central role in this analysis. We have gone to some length to identify these curves, and to explain what lies behind them. Nevertheless, we end this section of the chapter with a clarification that at first read may sound surprising: a firm's marginal cost curve may not be its supply curve and, in fact, a firm may not even have a supply curve. That said, price and output are still determinate, and the forces represented by supply and demand can continue to fill the function of Adam Smith's invisible hand.

The supply curve presented in this section of the chapter is the product of an abstract, stylized model of a competitive industry. Each firm in that industry is insignificant relative to the total of all firms, and the output of any one firm is indistinguishable from the output of any other firm. Being *fungible*, all of the firms' outputs can be pooled together when brought to market, and all units are sold at the same price. Accordingly, as we have said before, no individual firm has any economic power in the market, and the demand faced by each one of the firms is infinitely elastic at the price set in the market. This property of infinite elasticity is absolutely required for the firm's supply curve to be its marginal cost curve (starting, of course, at the minimum point on its average cost curve and rising above it). But unless the demand curve facing the firm is infinitely elastic, the very notion of a supply curve does not exist.

Why is this the case? We know that:

- The supply curve of the competitive firm is the locus of points that are traced out when the demand curve facing the firm shifts.
- All of the points that are traced out lie on the relevant part of the firm's marginal cost curve.
- The best profit optimality condition for the firm is to produce at a rate where $MC = MR$.
- Price and quantity are set by the intersection point where the firm's demand curve and marginal cost curve cross.
- For the $MC = MR$ condition to be satisfied where the *demand curve* intersects the marginal cost curve, the demand curve and the marginal revenue curve must be one and the same, for only then will $MC = MR = P$.
- For this to be true, the demand curve faced by the firm must be infinitely elastic.

End of story.

Infinitely elastic demand on the firm level is institutionally realistic for industries such as agriculture where a large number of firms are involved in the production of a fungible product. For instance, the red winter wheat produced by one farmer in Michigan is no different from the red winter wheat produced by any other farmer in Michigan. But in other industries, the output of one firm may be at least somewhat distinguishable from that of its competitors. For instance, small grocery, newspaper, and liquor stores in relatively close proximity to each other may compete with each other while retaining some individuality and thus some independent pricing ability. When they do, the demand curve for their products is not infinitely elastic. We show what this looks like in Exhibit 5.11.

The demand curve in the exhibit (which is labeled *d*) is sloping slightly downward. Thus the associated marginal revenue curve (which is labeled *MR*) is everywhere below the demand curve except at its intercept. This relationship follows from equation (5.10): with $MR = P\left(1 + \frac{1}{\eta_{X \cdot P}}\right)$, for $0 > \eta_{X \cdot P} > -\infty$, we have $MR < P$. The best profit output remains determined by the point where the *MR* curve intersects the *MC* curve. That point, which is labeled "A" in Exhibit 5.11, establishes $X^*$ as the best profit output. Going vertically up to the demand curve, we see that the unit price that the firm can charge for $X^*$ units is $P^*$.

The profit maximizing values of price and quantity, $P^*$ and $X^*$, define a point, labeled B, that is not on the *MC* curve. And the less elastic the demand curve is, the further point *B* is above point *A*. In general, the price, quantity relationship depends not only on the location of the demand curve

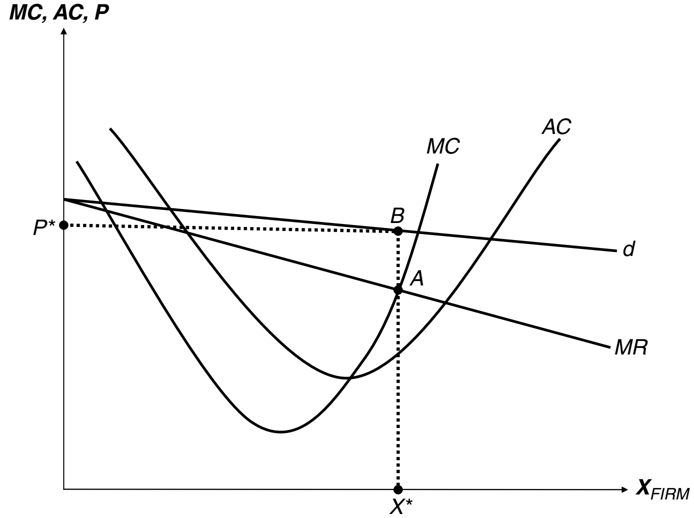


**EXHIBIT 5.11** The Optimal Output and Pricing Decision of a Firm Facing a Downward Sloping Demand Curve

labeled d, but also on the elasticity of that curve. If the elasticity changes as demand shifts, a tidy relationship between price and the quantity supplied is not traced out. Consequently, if a firm faces a demand curve that is downward sloping, that firm does not have a supply curve. Nevertheless, its output and price are, of course, determinate.

## APPLICATION TO A DEALER FIRM

Having considered the generic product *X*, we now focus on one specific industry and see how it operates within the broad context of our preceding discussion. We turn to the equity markets for that application.

### Some Basic Information about a Dealer Market

We considered four alternative structures for an equity market in Chapter 4: the quote-driven, dealer market; the continuous order-driven market; the periodic order-driven call auction; and the negotiated block market. Each of these structures requires a facility that must be owned and operated, and the operations could be described by a production function and cost curves. One of these, the quote driven, dealer market, provides a particularly clear and interesting representation of a firm that is supplying a service. This is the structure that we now focus on.

We discussed the dealer market at some length in Chapter 4. Here is a brief overview of the key points:

- Any dealer (say a used car dealer), both buys something (say a car) from a public customer who wants to sell the vehicle, and sells something (perhaps that car that the dealer has just bought) to a public customer who is looking to buy a vehicle. Equity dealers do the same thing. The "something" for them is shares of stock. Securities dealers who have an affirmative obligation to always make two-sided markets (that is, they must always be ready to both buy and sell) are referred to as "market makers." We use the terms "dealer" and "market maker" interchangeably.
- A dealer firm trades from its own inventory as a *principal*. When public investors want to buy, a dealer firm sells *from its own portfolio*, reducing a long position or going short. When public investors want to sell, the dealer firm buys *for its own portfolio*, reducing a short position or going long.
- Dealer firms indicate their willingness to buy or to sell by their quotes. A market maker's quotes consist of two parts: a bid to buy and an offer to sell, along with the share sizes that the two quotes are good for. Public investors sell at the market maker's bid and buy at the market maker's ask (which is sometimes called the *offer*).
- A dealer firm adjusts its quotes to the flow of public orders that it receives, continually seeking to find a level that best balances the forces of demand and supply. If the market maker's quotes are too low, the order flow it receives will be dominated by purchases at its offer. When this occurs, the dealer firm will soon acquire an unacceptably short inventory position, and the short position will lead it to raise its quotes. If the market maker's quotes are too high, the order flow it receives will be dominated by sells at its bid. When this occurs, the dealer firm will soon acquire an unacceptably long inventory position, and the long position will lead it to lower its quotes.

## Market Maker Services, Revenues, and Costs

We now turn to market maker operations. A market maker firm is characterized by the nature of the transactional services that it provides to investors, by the interesting way in which it receives its revenue, and by the unique costs that it incurs. We consider each of these in turn.

**Market Maker Services** The classic service supplied by a dealer is the provision of immediacy. Buyers and sellers arrive sporadically in a continuous trading environment, they need a way to meet, and they need prices to be stated that they can respond to. A dealer's quotes play this role. The market maker is continuously present, continuously quoting a two-sided market, buying when a public seller arrives and selling when a public buyer arrives. With market maker intervention, public participants can trade with immediacy even though they arrive at the market at different moments in time. The market maker is the medium through which public buyers and sellers effectively meet each other. While the earlier part of this chapter has focused on the undefined product, $X$, that is supplied by the competitive firm, we are now facing a more defined product, which is *immediacy*.

This is, however, a simplification. Market makers also offer additional services such as the supply of supplemental liquidity, which means the following. The market maker is not the ultimate source of liquidity for a market: public buyers provide liquidity to public sellers, and vice versa. At times, however, a demand/supply imbalance can develop and the market maker, by stepping forth with *supplemental liquidity*, can help to rectify that imbalance. In addition, market makers play an active role in finding the share prices that best reflect the broad market's desire to hold shares. But we do not consider these functions in this chapter. It pays to keep things simple. When it comes to the supply of market maker services, product $X$ for us is immediacy.

**Market Maker Revenues** How does a market maker firm charge for its provision of immediacy? A dealer firm (as distinct from a broker) generally does not charge a commission when it transacts with a customer. Rather, the firm realizes revenue from the spread between its bid and offer quotes, and the classic source of a dealer's profits is its bid-ask spread. Sometimes the spread is referred to as the *jobber's turn* ("jobber" is the traditional British name for a market maker). If the bid and the offer are reasonably stable over time, and if the dealer repetitively buys at a lower bid and sells at a higher ask, he or she profits from the round trips. A *round trip* is first buying and then selling (or, for a short position, first selling and then buying). The spread is the return for one round trip. The return for just one purchase or just one sale is conventionally taken to be half of the spread. From the market maker's customer's perspective, half of the spread is the per share price of trading with immediacy.

Trading volume given, wider spreads lead to greater profits for market makers. From a cost perspective, three factors determine the size of a dealer's spread: share price, trading volume, and the costs of market making. Share price is important because *percentage spreads* should be fairly constant across different price levels, all else constant. Look at it this way: A dealer firm by selling or buying 100 shares of a \$100 stock is supplying ten times the immediacy to a customer than it provides when buying or selling 100 shares of a $10 stock. Accordingly, the *dollar size* of the spread for the $100 stock will be ten times higher than the dollar spread for a $10 stock. This could, of course, be dealt with by working with percentage spreads instead of dollar spreads, as is commonly done in empirical analyses.

Taking the spread to be the source of a dealer's revenue, however, is a simplification. A dealer firm could charge a commission and some of them do. More importantly, a successful dealer has a sense of when to buy (because prices appear to be temporarily depressed) and when to sell (because prices appear to be temporarily elevated). Such buying and selling is referred to as *trading the order flow*. Successfully trading the order flow is another revenue source for a dealer. But for now we will ignore this one. Again, it pays to keep things simple.

**Market Maker Costs** Higher market making costs lead to wider spreads for the same reason that higher production costs generally lead to higher sales prices in any competitive industry. Two distinctive costs in particular characterize market maker operations, and these are the ones that we will focus on: the cost of carrying an unbalanced inventory, and the cost of trading with a better-informed counterparty (the cost of ignorance).

Regarding the *cost of an unbalanced inventory*, we know from financial economics (and Chapter 3) that an investor can manage risk by proper portfolio diversification, and that a properly diversified portfolio eliminates diversifiable risk and leaves the investor with an amount of *systematic risk* that the expected return appropriately compensates for. A market maker, however, is not a typical investor. A market maker is buying or selling not for his or her own investment purposes, but to supply shares to others who are looking to buy, or to absorb shares from others who want to sell. In so doing, the market maker is commonly in the position of holding a poorly diversified portfolio and, as a consequence, accepts risk that could have been diversified away. The expected return on a stock compensates the holder of the stock for accepting *nondiversifiable risk* (which in Chapter 3 we have also referred to as "systematic risk"). What compensates the market maker for accepting *diversifiable risk*? The answer is the bid-ask spread. Whatever makes the dealer's inventory control more difficult—be it information change, generally high price volatility, relatively infrequent order flow (particularly with a small-cap stock), or the stochastic nature of the order flow—translates into more costly market making and hence wider spreads.

The second cost that is endemic to a dealer firm, *the cost of ignorance*, is the cost to the firm of receiving an order from a better-informed trader. Say a market maker has posted a bid at 50 and an offer at 50.10 and that, without the market maker knowing it, the equilibrium price for the stock has jumped to 53. A better-informed trader comes in and buys at the 50.10 offer. The market maker loses from that trade because the bid and the offer will both rise as the news gets out more broadly, and the market maker will have sold at a lower value (50.10). After the market maker sees where prices are going, the firm will regret having traded with the better-informed counterparty. The principle is general: whenever a trade is triggered by a better-informed order hitting a market maker's quote, the market maker will suffer from something that is euphemistically called "ex post regret."

Of course, the dealer firm, like any firm, incurs other labor, capital, raw material, and administrative costs. We do not deal with these here except to note them. For the third time: it pays to keep things simple.

### A Dealer-Pricing Model

In this subsection we consider a simple dealer-pricing model that is based on the market buy and the market sell curves that a dealer firm faces. We see what is involved by assuming that just one dealer firm is the market maker for the stock. That is, for now, let the dealer be a monopoly market maker for the single stock that we are focusing on. Public investors buy from the dealer firm at the market maker's offer, and sell to the dealer firm at its bid. The market's desire to buy shares is described by a downward sloping function that has the properties of a demand curve; the market's desire to sell shares is described by an upward sloping function that has the properties of a supply curve. Following our discussion in Chapter 3, however, it is clear that there is only one demand curve in the market: it is the demand to hold shares. That is why we do not use the terms "demand" and "supply" here, but label the curves "buy" and "sell" instead. It is important to make this careful distinction.

Investor buy ($B$) and sell ($S$) curves are shown in Exhibit 5.12. If public investors could meet costlessly in a frictionless environment, the intersection of $B$ and $S$ would establish $P^*$ and $Q^*$ as the equilibrium values of share price and the number of shares exchanged. But buyers and sellers do not arrive at the market at the same time, and it is not a simple matter for them to find each other in time. They need a mechanism that enables their orders to meet so that executions may be realized. In an order-driven market, the limit-order book is such a mechanism. In a quote-driven market, the dealer is such a mechanism. We focus here on the latter. For now, assume that just one, single, monopoly dealer is making the market and supplying immediacy, the service that the public investors are demanding.[3]

---

[3] This discussion follows Harold Demsetz, "The Cost of Transacting," *Quarterly Journal of Economics* 82, no. 1 (1968), 33–53.

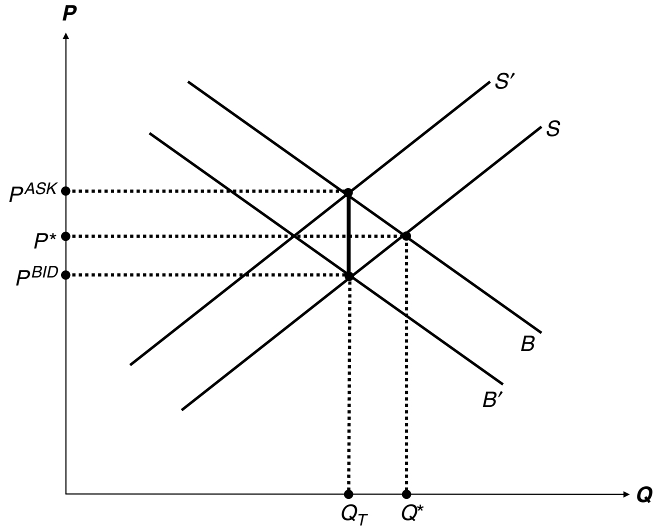


**EXHIBIT 5.12** A Dealer Firm's Bid and Ask Prices, Transaction Volume, and the Supply of Immediacy

The market maker satisfies the public's demand for immediacy by continuously standing ready to buy and to sell. But the service is not free; the dealer sells to buyers at higher prices and buys from sellers at lower prices. The curves labeled $S'$ and $B'$ in Exhibit 5.12 reflect this reality. $S'$ is the supply curve of immediately available shares; $B'$ is the demand curve for shares that can be immediately bought. Together, $S'$ and $B'$ reflect the dealer's sale of immediacy. This is how it works.

At each quantity, the vertical distance between $S$ and $S'$ and between $B$ and $B'$ shows the price imposed by the dealer. The dealer's charge is very much like a sales tax, except that a specific service is being directly supplied to the customer who pays this charge (whereas the receipts from a sales tax simply go to the government), and this monopoly dealer model looks very much like the sales tax model in Chapter 3 [flip back to Exhibit 3.10(a)]. For simplicity, let the vertical distance between the primed and unprimed curves be the same at all values of $Q$ and for both sets of curves. Equilibrium is given in Exhibit 5.12 by the intersection of $S'$ and $B$ and, equivalently, by the intersection of $S$ and $B'$. The intersection of $S'$ and $B$ establishes the market ask ($P^{ASK}$) and the intersection of $B'$ and $S$ establishes the market bid ($P^{BID}$). The transaction volume at both the bid and the ask is $Q_T$. This equality keeps the dealer's inventory in balance.

We thus have the following. Buyers and sellers arrive sporadically at the market and generally do not find each other *in time*. Traders cannot leave orders in the market because no mechanism exists for handling them. Therefore, despite the fact that a frictionless market would clear at $P^*$ with $Q^*$ shares transacting, the real-world market, by itself, cannot operate. The market forms because a dealer posts an ask of $P^{ASK}$ and, at this price, sells $Q_T$ shares to the market (given the buy curve labeled $B$, $Q_T$ is the number of shares that the market wishes to purchase at $P^{ASK}$). The dealer also posts a bid of $P^{BID}$ and, at this price, buys $Q_T$ shares from the market (given the sell curve labeled $S$, this is the number of shares that the market wishes to sell at $P^{BID}$).

Accordingly, $Q_T$ shares pass via the dealer from public sellers to public buyers. The dealer has accommodated the public's demand for immediacy by selling from its inventory and by buying for its inventory. The dealer is compensated for providing this service by buying at a relatively low price (its bid) and by selling at a relatively high price (its ask). Note that, because transacting in the market is not costless, the actual number of shares traded, $Q_T$, is less than the frictionless market transaction value, $Q^*$.

## Dynamic Dealer Pricing

The dealer-pricing model that we looked at in the previous subsection is a good framework for understanding how the market maker facilitates buyers and sellers meeting one another. But it is just the beginning. It is a monopoly dealer model while dealer markets are generally competitive. Most notably, it does not incorporate uncertainty, and it simply posits that a fixed vertical distance separates the $S'$ and $S$ curves, and the $D'$ and $D$ curves. The vertical distance produces a bid-ask spread and, for a public investor who wants to buy or to sell, half of the spread is the price of immediacy. But what determines the size of the spread and thus the cost of making a one-way (as distinct from a round trip) transaction? Dealer quotes are not static during a trading day; what affects their dynamic behavior? As we address these issues, two considerations are paramount: (1) the costs of market making and (2) the strength of interdealer competition. We are dealing with costs in this chapter and we turn to competition in the chapter that follows.

We have noted two costs in particular that characterize dealer operations: the cost of carrying an unbalanced inventory and the cost of transacting with a better-informed investor (the cost of ignorance). Regarding the cost of an unbalanced inventory, let the dealer firm start a trading session with a desired inventory of zero shares. If by chance the early flow of orders from public investors is dominated by buys, the market maker will soon acquire a short position; the firm will attempt to return to a flat position (and, in so doing, to re-liquify) by raising its bid (to encourage more public sells) and by raising its offer (to discourage more public buys). If by chance the early flow of orders is dominated by sells, the market maker will acquire a long position, and the firm will do just the opposite (lower its bid to discourage sells and lower its offer to encourage buys).

A dynamic pricing policy also addresses the second major cost faced by a market maker firm: the cost of ignorance. Let a couple of public orders arrive to buy at the dealer's ask. The dealer does not know whether these orders are motivated by reasons unique to individual traders or by news that the dealer does not yet know about, but which some public traders do. And so these orders convey a noisy signal. The arrival of a further buy order in a relatively brief interval of time would strengthen the signal. Accordingly, the dealer firm's optimal selling price (ask) for further buy orders is higher and, given the possibility of the buyers really showing up, the dealer firm increases its ask. In so doing, it maintains optimal protection against the better-informed buyers. Similarly, the dealer firm decreases its bid after a succession of sell orders that could have been motivated by bearish news that have arrived in the market in a relatively brief interval of time.

Therefore, in relation to the costs of both risk bearing and ignorance, the dealer firm increases its quotes in response to buying pressure from the public, and decreases its quotes in response to selling pressure from the public. This dynamic pricing policy has been modeled in the dealer pricing literature. The approach taken is to assume that the arrival of orders and the behavior of an underlying clearing price can be described by statistical processes, and that the dealer makes pricing decisions in relation to the anticipated order flow in a multiperiod context. In the theoretical literature on dealer pricing, the optimal value of the market maker's quotes at any current moment in time is given by the solution to a *dynamic programming model*.

Dynamic programming reasons in reverse. It first considers the dealer's wealth as of some future point in time and the quotes at the final decision point that will maximize the expected utility of this wealth. The dynamic programming model then considers the second-to-last decision point and solves for the quotes that will put the dealer in the best possible position as of the last decision point. Continuing to work backward through time, the dynamic program finally solves for the optimal value of the quotes as of the starting decision point. This solution satisfies the ultimate objective—maximization of the expected utility of final wealth.

And so, according to the theoretical, dynamic dealer-pricing models, at any moment the dealer sets its current quotes and sees what happens. As trading occurs with the passage of time, the dealer firm keeps remaking its pricing decisions, always considering how the situation may look as of the close of the trading day, and then working backward to the current moment to determine how best to position its bid and offer in light of what might subsequently come to pass.

Do real-world dealers actually operate this way? Few of them probably know about dynamic programming models but, nevertheless, they can act as if they do, and the models yield insightful implications.

## Dynamic Inventory Control

Hand in hand with the dealer firm's dynamic pricing policy is its inventory policy. The dynamic pricing policy is motivated by a desire to maintain reasonable bounds on inventory fluctuations. To see what is involved, we now focus on the dealer firm as it services the receipt of buy and sell orders. As predominantly buy orders arrive, the dealer firm sells shares and acquires an increasingly large short inventory position; as predominantly sell orders arrive, the dealer firm buys shares and acquires an increasingly large long inventory position. Even if the bid and ask quotes remain fixed and probabilistically balanced, the dealer's inventory, by the law of large numbers, will fluctuate between ever-increasing bounds (even though the proportion of buys versus sells may get ever closer to 0.5, the absolute difference between them will on expectation get ever larger). The inventory fluctuations would eventually result in a portfolio imbalance (either long or short) that would force the dealer firm into bankruptcy.

The dealer firm controls its inventory by placing bounds on the long and short positions that it will accept and, as these bounds are approached, the dealer firm adjusts its quotes to reduce the probability that further moves away from the desired inventory position will occur. The change of the quotes as an inventory control mechanism is illustrated in Exhibit 5.13 where price is shown on the vertical axis and the dealer's inventory position is on the horizontal axis. The negatively inclined curve labeled *Ask* in the figure shows how the dealer's asking price changes with its inventory position; the negatively inclined curve labeled *Bid* shows how the dealer's bid price changes with its inventory position. The ask and bid both decrease as the dealer's inventory increases above the desired value of zero, and the quotes both increase as the inventory falls below its desired value of zero. As the location of the spread shifts, its size is kept fixed (the linear Ask and Bid curves are parallel). A constant dollar spread is institutionally realistic and it is a useful analytical device.

The tighter the bounds placed on allowable inventory fluctuations, the steeper will be the ask and bid curves in Exhibit 5.13, and the larger will be the spread at each inventory level. Tighter inventory bounds imply less willingness of the dealer firm to supply marketability services; looser inventory bounds imply just the opposite. Therefore, the slope of the two curves and the vertical distance between them reflect the willingness of a dealer firm to make a market in a stock. A dealer firm that is willing to commit more

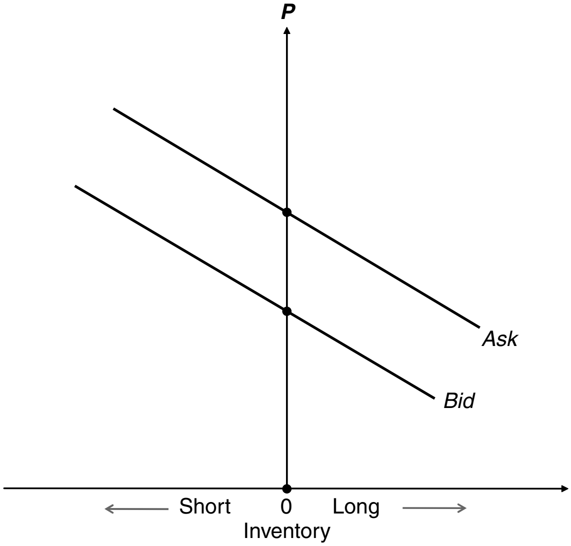


**EXHIBIT 5.13** A Dealer Firm's Quotes and Its Long/Short Inventory Position

capital to market making makes a better, more liquid, more orderly market. Note that with flatter curves and smaller spreads, transaction prices are less volatile.

### The Competitive Structure of a Dealer Market

In actual markets, the number of equity market makers can range from just a couple for thin, infrequently traded stocks, to fifty or more for large cap, frequently traded stocks. Each dealer firm, in reality, controls its own quotes and different firms typically set different bid and offer prices. We abstract from this. Let the number of firms be large enough so that we can ignore the impact that any individual firm's pricing decision might have on *the* market price of immediacy (*the* bid-ask spread set in the market that customers all see and respond to). To facilitate the discussion, also assume that the immediacy provided by each and every dealer firm is fungible. These assumptions are standard for a competitive market where all participants are price takers, not price setters.

Here is the rest of our model structure. The aggregate order flow for a specific stock to the competitive dealer market is negatively related to the *common price* charged for immediacy (the individual dealer spread that is the same for all dealers). Wider spreads increase the cost to the public of buying or of selling shares either for their own liquidity reasons or because their own assessments of the stock's value have changed. Wider spreads also decrease the flow of informed orders because news trading is profitable only when the price change that a news bit implies is greater than the cost of trading (which we are taking to be the bid-ask spread). Let's reword this one. Because a news event leads to profitable trading only if its impact on price is greater than the cost of trading, wider spreads make trading unprofitable for a larger array of news events.

Consequently, the industry faces a downward sloping demand curve for its provision of immediacy. This downward sloping curve, when interfaced with the supply of dealer services, determines an equilibrium spread and transaction rate. However, to obtain the supply curve of the competitive dealer industry, some issues concerning dealer output and cost control have to be addressed. We do this next.

## Dealer Output and Cost Control

We have just seen that, for all dealers, the size of the bid-ask spread faced by customers controls the rate at which the competitive dealers in aggregate transact with the public (with transaction rates varying inversely with the size of the spread). Each particular dealer firm has a way to control its own supply of immediacy—it sets position limits that force it to stop buying when an upper inventory limit has been hit, and to halt selling when a lower inventory limit has been hit. The wider these bounds, the more transactions an individual dealer firm will expect to handle for a given size of the bid-ask spread.

But just as the dealer firm does not know exactly what its output will be, neither does it know exactly what its costs will be. For instance, tightening its inventory bounds limits the number of transactions the firm can participate in and, in so doing, reduces its *expected number* of transactions; however, if buys and sells happen, by chance, to be nicely interspersed, a dealer firm can enjoy a larger transaction rate without its inventory bounds being hit even if they are relatively tight. On the other hand, if buys happen to be followed by more buys, and sells happen to be followed by more sells, the inventory bounds will be hit more frequently even if they are relatively wide.

As buy and sell orders arrive in response to liquidity trading and news arrivals, the market for a stock changes and the level of the quotes adjusts. The quote adjustments offset the revenue the dealer firms get from the spread. Say a string of buys comes into the market and the quotes are pushed up. Perhaps, as this happens, a dealer firm first sells at its offer; then, as the quotes rise in the market, the dealer firm's new bid is either (a) *closer to*, or (b) *above* the offer it has just sold at. What does this mean? The dealer, in making a round trip, either does not profit from the full spread [this is (a) above, the new bid is closer to but still below the old offer] or does not even realize a positive spread [this is (b) above, the new bid is *above* the old offer]. The bottom line on this one is that dealers hate trending markets (a trend can cut deeply into their profits), but love reversing markets (it is the reversals that enable the dealers to realize their profits). On the other hand, if a dealer buys just before a market turns up (or sells just before it turns down), its realized round trip profits will be greater than the spread.

The dealers' cost curves reflect the differences between quoted spreads and realized spreads. Realized spreads that are less than quoted spreads (or that are negative) add to the dealer firm's operating costs, and realized spreads that are greater than the quoted spreads offset some of the dealer firm's operating costs. Because the deviation of realized spreads from quoted spreads is stochastic, we analyze a dealer firm's *expected* average and *expected* marginal cost curves.

You can view it this way. The dealer firms respond to the spread that is collectively set (and is plotted on the vertical axis). They make their output decisions with respect to the price of immediacy and their expected costs curves. They control their expected output rate (which is plotted on the horizontal axis) by their position limits. But the vagaries of the order flow impact their realized output rates and their realized operating costs. We expand further on what this means in the subsection that follows.

### Dealer Cost Curves

We now turn to the dealer firm's cost curves. Unlike the deterministic cost curves for product $X$ shown earlier in this chapter, the dealer cost curves reflect the existence of risk. We assume that the individual dealer firm takes account of the price that it will receive for its services (the size of its bid-ask spread), along with its expected average cost, expected marginal cost, and the expected number of transactions that it will realize in a given period of time (e.g., a trading day). In dealing with these curves, we focus on the long-run cost relationships and thereby avoid the need to distinguish between fixed and variable costs.

The dealer firm's expected average cost, $E(AC)$, and its expected marginal cost, $E(MC)$, are shown graphically in Exhibit 5.14. Price (the bid-ask spread) is on the vertical axis, and the expected number of transactions per unit of time is on the horizontal axis. In the earlier subsection, "A Dealer-Pricing Model," we considered a monopoly dealer firm, and U-shaped cost curves of the kind shown in Exhibit 5.14 could apply to the monopoly

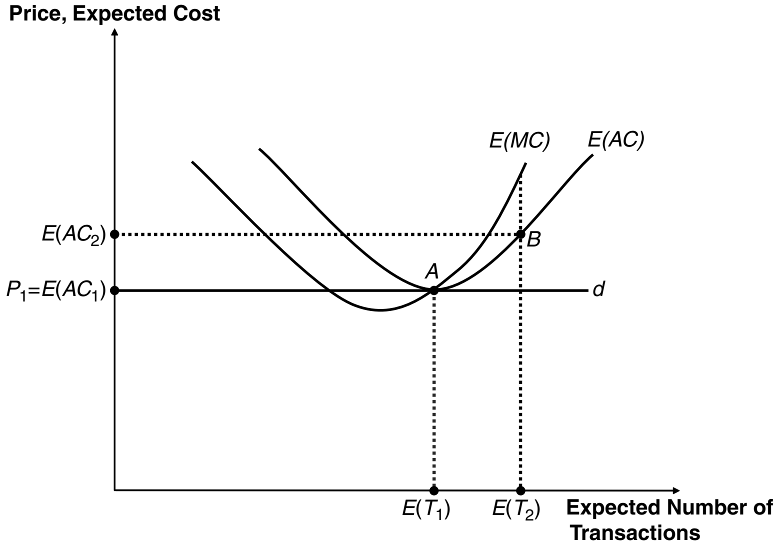


**EXHIBIT 5.14** A Dealer Firm's U-Shaped Expected Average and Marginal Cost Curves

dealer. But we now turn our attention to a competitive dealer firm so that we can bring the firm's supply curve back into the picture.

In Exhibit 5.14, the horizontal line with an intercept at $P_1$ is the infinitely elastic demand curve faced by the competitive dealer firm. The figure also shows the expected number of transactions [$E(T_1)$, the expected output rate] of a competitive dealer firm when the bid-ask spread is $P_1$. We have located $P_1$ with reference to point $A$, the specific point where the dealer firm's best profit output is at the minimum point on its expected average cost curve.

Because order arrival is stochastic, the competitive dealer firm does not know *ex ante* how many transactions per trading session it will actually participate in at the price $P_1$. Rather, given $P_1$, the firm adopts an inventory policy that implies an expected number of transactions of $E(T_1)$. That is, as we discussed above, the dealer firm places bounds on the magnitude of the inventory fluctuations that it will accept. Knowing that the location of its quotes will change as large long or short inventory positions are acquired and, given its inventory control policy, the representative competitive dealer firm expects a transaction rate of $E(T_1)$. The firm could, however, increase its expected transaction rate by widening the bounds on its inventory positions. For instance, consider point B in Exhibit 5.14. If the firm sets bounds that are wider than those associated with $E(T_1)$, its expected number of transactions would be a value such as $E(T_2) > E(T_1)$. Or, if it sets tighter bounds, its expected number of transactions would be less than $E(T_1)$.

*Ex post*, the dealer firm will have realized an actual transaction rate and will have experienced a specific pattern of inventory adjustments. If by chance buy and sell orders have arrived in a well-interspersed pattern, the inventory fluctuations required to service the order flow will have been small and the dealer firm will have participated in more than $E(T_1)$ transactions. Or if, by chance, the buy and sell orders have arrived in clusters (either because of the vagaries of order arrival or because of the pattern of adjustment to information change), the inventory fluctuations required to service the order flow will have been large, inventory bounds will have been hit more frequently, the dealer will have participated in fewer than $E(T_1)$ transactions, and its realized costs will have been higher. Nevertheless, the firm sets the maximum inventory limits *ex ante* and, having set them, anticipates that it will participate in $E(T_1)$ transactions at an average cost of $E(AC_1)$ during the next trading session.

The cost curves shown in Exhibit 5.14 have the traditional U-shaped pattern. The decrease in expected average cost over outputs ranging up to $E(T_1)$ is reasonable for the dealer firm. Regardless of the transaction volume it may sustain, a firm can make a market in a stock only if it is willing to commit at least a minimum amount of labor and financial capital to posting quotes, to staying adequately informed about the company, and to gaining an understanding of the dynamic behavior of the order flow for the security. These expenditures do not increase proportionately as the firm expands its market making activities in the stock, and so economies of scale are realized up to an expected transaction rate of $E(T_1)$.

But as the dealer firm's market making activities in a stock continue to increase, it must accept ever-greater inventory fluctuations in its trading portfolio. As these fluctuations get larger, they become increasingly risky and therefore costly, and diseconomies of scale set in for expected volumes greater than $E(T_1)$. Thus the dealers expected average cost curve is U-shaped.

### Externalities and the Supply of Dealer Services

One other cost factor affects the output of competitive dealer firms and thus the industry's supply curve: economies of scale that are external to the individual dealer firms but that are internal to the industry. Scale economies may be realized as aggregate dealer trading increases for a security for the following three reasons: (1) as the size of an issue grows, the flow of information about it that comes to the market increases (the corporation itself releases more information and a larger number of financial analysts pay attention to the stock); (2) as the aggregate order flow increases, dealers can more quickly work off an unbalanced inventory position, both by trading with the public at an increased rate and by trading with each other (an action that is called *interdealer trading*); and (3) as the aggregate market for a stock increases (all else equal), the amount of liquidity trading will likely increase in relation to information trading, which reduces the risk associated with any given inventory imbalance. All told, more frequent trading makes it easier for a market maker to reliquify (i.e., to get out of an accumulated position, either long or short), and being able to trade down faster to a comfortable position reduces a market maker's risk exposure to information change. And so, larger trading volumes result, all else equal, in tighter spreads.

The derivation of the industry supply curve under conditions of scale economies that are external to the individual dealer firms follows the methodology described with regard to the generic product $X$ in this chapter's earlier subsection, "Externalities and the Supply Curve." Exhibit 5.15 shows the industry's long-run supply curve of dealer services under these conditions. It is the curve labeled $S$ in Exhibit 5.15(a). When the market demand for dealer services is given by the curve labeled $D_1$, the price of immediacy is set at $P_1^*$ where $D_1$ crosses $S$. At this point, firms in aggregate operate at an expected total transaction rate of $E(T_{A1}^*)$ *and [look at Exhibit 5.15(b)] the average cost curve for the* $i^{th}$ individual firm is the higher curve labeled $E(AC_{i1})$. *Firm* $i$'s best profit output at the price $P_1^*$ is the expected output $E(T_{i1}^*)$, and $E(T_{i1}^*)$ times the number of firms in the industry ($N$) is $E(T_{A1}^*)$. Let's repeat. The consistency condition underlying this solution is that, when the expected aggregate number of market transactions is $E(T_{A1}^*)$, the expected average cost curve for the representative firm is the curve labeled $E(AC_{i1})$, the best profit output for that firm is $E(T_{i1}^*)$, and $N[E(T_{i1}^*)]$ in fact equals $E(T_{A1}^*)$ (remember, $N$ is the number of dealer firms in the competitive industry).

If the market demand shifts to the curve labeled $D_2$, equilibrium is again achieved for the industry and for each of the individual dealer firms that it comprises. This new solution gives an expected aggregate transaction rate of $E(T_{A2}^*) > E(T_{A1}^*)$. Because the expected transaction rate is greater and given the economies of scale that are external to the individual dealer firms, the expected average cost curve for each firm is lower, and its minimum point is further to the right [as shown in Exhibit 5.15(b) by the relationship between $E(AC_{i1})$ and $E(AC_{i2})$]. Because the $E(AC_{i2})$ curve is lower, the market clearing price that would lead the firm to produce at the minimum point on its expected average cost curve is also lower, it is $P_2^* < P_1^*$. Because the minimum point on the $i^{th}$ firm's average cost curve is further to the right, we have $E(T_{i2}^*) > E(T_{i1}^*)$ and so, even if the number of firms in the

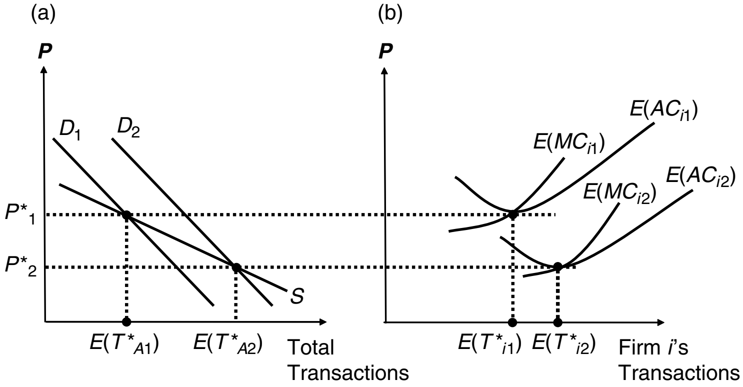


**EXHIBIT 5.15** An Industry's Supply of Transaction Services [5.15(a)] and the Expected Cost Curves of an Individual Dealer Firm [5.15(b)] under Economies of Scale That Are External to the Individual Firms but Internal to the Industry

industry remains fixed, aggregate industry output increases. In the figure shown in Exhibit 5.15(a), the supply curve for the competitive industry is downward sloping (which, as we noted in an earlier subsection of this chapter, could happen).

The downward sloping supply curve for a competitive dealer industry is of particular interest. It suggests that we should expect to observe tighter competitive dealer spreads in larger markets, and we do.

## FROM GENERAL PRINCIPLES TO THE DETAILS OF A SPECIFIC MARKET

In the first part of this chapter we laid out a basic structure for analyzing the supply propensities of an industry. We started with the production function, proceeded to cost curves, and ended with firm and industry supply curves that could be matched with a demand curve to establish price and quantity for the industry, and the outputs of the individual firms in the industry. The analysis shows in a simplified setting how a micro market may operate from a production and supply perspective, as well as from a demand point of view.

How applicable is the analysis to real-world markets? The broad principles that it establishes are definitely useful. We see how inputs into production are optimized, and how best profit outputs can be achieved. We see that, in a competitive environment, firms have the incentive to produce at rates that equate their marginal costs with the prices of their products. This is of particular interest. We saw in Chapter 3, in a two-good model ($X$ and $Y$) that, when a household maximizes its utility, the ratio of prices for the two goods equals the ratio of their marginal utilities. We now see that the price of each good equals its marginal cost. Pulling it all together delivers a powerful result: If two industries ($X$ and $Y$) are competitive and if long-run equilibrium is attained, the ratio of the marginal utilities for each and every consumer will equal the ratio of the marginal costs of producing the two goods $\left(\text{that is, } \frac{MU_X}{MU_Y} = \frac{MC_X}{MC_Y}\right)$. Equality on the margin means that, collectively, the economy is getting the most it possibly can from its scarce resources. What more could be asked of a micro market?

General principles covered in the earlier part of this chapter are applied in a more specific setting in the latter part of the chapter, where we zeroed in on the market for equity dealer services. Several things can be learned from this exercise. The general analytical framework can be usefully applied, but it had to be further developed. Questions had to be addressed concerning just what the service is that is being supplied by the industry; just how "price" is defined for this industry; and just what the unique costs are that the firms in this industry incur. For the dealer firm, the service is not shares offered or purchased per se, but a more abstract product called "immediacy"; price is not an explicit dollar fee, but a bid–ask spread; and the two big costs are not the costs of labor and of physical capital, but the cost of carrying an unbalanced inventory and of the cost trading with a better-informed customer. To handle these costs, risk had to be introduced into the analysis.

This is our concluding remark. There is a method and a methodology for studying a micro market, but each market has its own unique characteristics. Thus thought has to be given to how the tools of microeconomic analysis can best be applied to any particular, real-world market. This is all part of the fascination of the subject.

## CHAPTER SUMMARY

In this chapter our focus has switched from the consumer side of the market to the producer side. We have considered production functions, cost curves, and supply curves (their counterparts from the consumer side are utility functions and demand curves). Having analyzed both the demand and the supply sides of the market, we are in a good position to investigate how competition plays out in a marketplace, the topic that we turn to in Chapter 6. Here are the highlights of Chapter 5:

- The starting point for supply analysis is a firm's production function, a technological description of how the factors of production that a firm uses can be combined in varying proportions to produce the good or service that it is supplying to the market. We first focused on our generic product, *X*. We simplified the discussion with no substantive loss of generality by dealing with just two productive factors, which we called "labor" and "capital."

- Two sets of curves were defined: *isoquants* (each isoquant is the locus of all of the different labor and capital combinations that give the same output), and *isocost lines* (each isocost line is the locus of all of the different labor and capital combinations that require the same total expenditure). A firm makes its input and output decisions with regard to a family of isoquant and isocost lines. Wherever an isoquant is tangent to an isocost line, total production costs are minimized for that output and, reciprocally, total output is maximized for that cost. Accordingly, the tangency solution satisfies one requirement for profit maximization: the least cost combination of inputs condition.
- The set of isoquant, isocost line tangency points defines an upward-sloping line that is called the firm's expansion path. If equal movements out along a linear expansion path (attributable to constant incremental increases in the inputs of labor and capital) result in increasing increments of output, the firm is operating under conditions of increasing returns to scale. If constant increments of output are realized for equal movements along the path, the firm is operating under conditions of constant returns to scale. If decreasing increments of output are being realized, the firm is experiencing decreasing returns to scale.
- Total cost is not a constraint for a firm in the way that a budget constraint is for a household. If a firm can use more inputs and produce a larger quantity profitably, it can operate at a higher level of expenditure. The equilibrium output level for a firm depends on the price that it can sell its product at, and on its production costs. The information needed to construct the firm's cost curves can be extracted from its isoquant, isocost line mapping.
- We obtained the firm's long-run and short-run total cost curves. These were used to get the average and marginal cost curves. In doing so, we distinguished between fixed costs and variable costs.
- With the firm's marginal and average cost curves in hand, we identified the second requirement for profit maximization: the firm's best profit output. Given that the least cost combination of inputs condition is satisfied, the firm's best profit output is the output level that equates marginal cost with marginal revenue, subject to average variable costs being less than average revenue (recall that for the competitive firm, average revenue is the product's price).

With the firm's marginal cost curve and best profit output identified, we established that the competitive firm's supply curve is its marginal cost curve above the minimum point of its average cost curve. The analysis was then extended in two ways. First, we showed how the supply curve for the industry is affected by economies of scale that are external to the individual firms but internal to the industry in aggregate. Second, we clarified that the marginal cost curve is the supply curve of the firm only for a perfectly competitive industry. You will appreciate as you read on that both of these extensions are important (the first was applied in the latter part of this chapter, and the second is germane for the next chapter, Chapter 6).

- Except for a concluding section, the remainder of the chapter considered how our cost and supply analysis applies to a dealer market. After providing some basic background information on this market, we identified dealer services (first and foremost, the provision of immediacy), dealer revenues (the classic source is the dealer's bid-ask spread), and dealer costs (the cost of carrying an unbalanced inventory and the cost of trading with a better informed investor). Each of these items (services, revenues, and costs) is unique to the dealer firms. More general economic analysis helps to understand them, but each requires special thought.
- As you no doubt have noticed, microeconomic analysis typically involves the search for equilibria, and this is what we did with regard to the dealer market. We first presented a simple pricing model for a monopoly dealer, and then considered a dynamic dealer pricing model.
- Hand in hand with the dealer firm's dynamic pricing policy is its need to maintain reasonable bounds on its inventory fluctuations. With the firm's inventory policy depicted, we turned to a competitive dealer market and obtained a set of representative average and marginal cost curves for a competitive dealer firm. This derivation required some special thought, as the cost curves were not obtained from a dealer firm's production function. Nevertheless, the standard U-shaped average and marginal cost structures held for the competitive dealer firm. We made one adjustment, however. Rather than referring simply to output and costs, we operated in the space of the *expected* number of transactions and *expected* costs.
- As noted, we have shown earlier in the chapter how the supply curve for an industry is affected by economies of scale that are external to the individual firms in the industry, but are internal to the industry in aggregate. We applied this analysis to the industry's aggregate supply of dealer services. Economies are realized as the market for an individual stock grows—that is, as the total value of shares outstanding for an

issue (referred to as the issue's *cap size*) increases: more information is available about larger cap firms, their stocks' trading volume is greater, they trade more frequently, and the ratio of liquidity traders to information traders may be higher. All of this facilitates inventory management and cost control for the individual firms and, in so doing, shifts the firms' cost curves down, and increases the elasticity of the industry's aggregate supply of dealer services.

- Having investigated how microeconomic analysis can be applied to the dealer market, we concluded the chapter with the following thought: "There is a method and a methodology for studying a micro market, but each market has its own unique characteristics. Thus thought has to be given to how the tools of microeconomic analysis can best be applied to any particular, real-world market. This is all part of the fascination of the subject."

## LET'S CONSIDER AN APPLICATION: THE OPTIMAL SCHOOLING BEHAVIOR OF FISH

We have devoted considerable attention in this chapter to understanding the supply behavior of firms in a perfectly competitive industry. You might question the extent to which actual firms in real-world, competitive industries behave in accordance with our model. Of course the owners of the firms respond to economic variables. Of course they will strive to bring more to market when market prices are higher and/or when input prices are lower. But do they really know their cost curves? Do they really *maximize* their profits?

Of course not. Information is never perfect enough for anyone to be able to do this, and decision making is never sharp enough to deliver perfect world answers. But, as we pointed out in Chapter 2, our microeconomic models do not claim to describe the behavior of every economic agent. Rather, they seek to predict what the representative household or firm will do (a positive as distinct from a normative application). We describe the central tendency of a group and, in so doing, are actually modeling the behavior of the group as an aggregate. An intriguing question to ask in light of this is how much explicit rational thinking is actually required on the part of individual decision makers for our predictions of group behavior to be reasonably accurate? This inquiry is what our thought question for this chapter is all about.

Let's consider a species that is not known for having much grey matter—fish. An interesting thing about many types of fish is their schooling behavior. There are benefits to schooling. It is believed that schooling may reduce friction and so enable energy to be conserved when the fish are on the move. There can be safety in numbers as some foolish predator may think that 75 little fish, all packed tightly together, are actually one big fish. There is a confusion factor: The mass behavior of a sizable school can confuse a predator (think of the difficulty that some of us have catching just one tennis ball when two or more tennis balls are thrown at us at the same time).

Schooling also allows individual fish to communicate rapidly with one another. Have you ever observed a school of little fish? Have you noticed how, with lightning speed, the school can change course and set off in a new direction? Eyesight and highly sensitive lateral lines on the fishes' bodies enable them to do this. Here is the picture. A predator suddenly appears, and the fish closest to the predator detect the danger, change direction, and rapidly dart away. Knowledge of this maneuver is transmitted virtually instantly throughout the entire school, which, as a group (perhaps minus one or two) dashes away.

Animal scientists have asked the question, "Why do fish school?" and they have advanced testable hypotheses such as those we have just presented. An economist will ask another question or two. Is there an *optimal* configuration for a school? Why might the optimal configuration be different for different types of fish (an exercise in comparative statics)? And there is the classic economic question: Are there trade-offs involved that would enable us to model the optimal configuration of a school for a given kind of fish? With these economist-type questions in mind, we ask you to work your way through the following thoughts and questions:

1. Using methodology that you are now familiar with (and may appreciate for the simplicity that it delivers), introduce a trade-off situation for the fish in a two-factor model. To start, we can pick one of the benefits of schooling that we have just identified. Let's select the transmission of an information factor, and assume that information can be transmitted more quickly from fish to fish if they are schooled more tightly together. Now, to have a trade-off situation, something must be given up if the school is tighter. That is why a second factor is needed. Can you think of a suitable second factor?
2. Well, here is a suggestion for the second factor: A school of fish moving though a food field will be able to consume more food

(*Continued*)

(*Continued*)

per fish when the fish are less tightly grouped together. So, there you have it. The trade-off is between protection from predators (which is greater the shorter the distance between the fish), and per fish consumption (which is greater the greater is the distance between them). We ask you to plot this trade-off. So that we can be in synch on this, put protection ($P$) on the vertical axis and consumption ($C$) on the horizontal axis.

3. Hopefully you have made your line downward sloping (after all, we are dealing with a trade-off). The line plays the role of a technologically determined efficient frontier (a kind of budget constraint). Our question now is, what must change in order for there to be a movement along the curve? That is, what variable, by changing, will move the fish from one point on the frontier to another point?
4. Okay, question 3 was fairly easy. The distance between the fish is the variable that we were asking about. When the school is more spread out it realizes more $C$ and less $P$, and when it is packed more tightly together it enjoys more $P$ and less $C$. So let's proceed to a more challenging question. Is the efficient frontier linear or curved? If curved, is it concave or convex from below?
5. The answer to number 4 depends on the marginal productivity of the distance between the fish in delivering $P$ and $C$. Call the variable "distance." Assume that increasing distance has diminishing productivity in terms of delivering $C$. Further assume that decreasing distance has diminishing productivity in terms of delivering $P$. Now answer the question in number 4.
6. With the efficient frontier (a.k.a. the technologically determined budget constraint) identified, how is an optimal point along it found? By, of course, interfacing the frontier with a family of indifference curves. Accordingly, we ask you to add a family of indifference curves to your graph. But do not simply draw these curves by rote. Make sure that their curvature makes sense to you.
7. Can the model be extended to suggest why optimal schooling behavior is different for different kinds of fish? Can we formulate some comparative statics? Of course we can. We ask you to give it a try. Contrast the optimal solution for two schools of fish, A and B, with the fish in school A being smaller than the fish in school B. Assume that, in the limit, for both schools, as the distance between the fish becomes infinitesimally small, protection reaches the same maximum ($P^{MAX}$). As the distance between the

fish increases, consumption ($C$) increases for both schools, but assume that this too reaches a limit. Let the limit value of $C$ be greater for the larger fish than it is for the smaller fish ($C^{MAX(\mathrm{A})} < C^{MAX(\mathrm{B})}$). Now introduce their indifference curves and contrast the resulting equilibrium solutions. Which type of fish does your model suggest will school more closely together, A or B?

8. We started this application by asking how much explicitly rational thinking is required on the part of individual participants, in any kind of market, for our predictions of group behavior to be reasonably accurate. What do you think? Do fish behave as if they know all of the things that you have just figured out? And here is one final question: Do fish really have indifference curves?

## TraderEx SIMULATION: WHAT IS IT LIKE TO BE A SECURITIES DEALER?

Again we turn to the TraderEx simulation software that is available in *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course* to give you a further opportunity to work with some of the concepts that we have developed in this text. In this chapter we have described the challenge a dealer firm faces keeping its inventory of shares within reasonable balance. The dealer firm's main control mechanism is its quote setting: it will raise its quotes to work off an excessively large short position, and it will lower its quotes to work off an excessively large long position. The dealer firm's second control mechanism is to trade with other dealers: it will buy in the interdealer market to reliquify after having acquired an unduly large short position, and it will sell in the interdealer market to reliquify after having acquired an unduly large long position. We have also seen that a dealer firm that is willing to accept larger inventory swings (a) will have a larger expected trading volume and (b) will operate at a higher level of expected costs. In this simulation exercise, we would like you to experience these realities by trading as a dealer in our simulated environment. Here is what you can do:

- You will start each simulation run with zero cash and zero shares of the risky asset for which you will be the market maker.
- There are no explicit trading costs in the simulation. You can borrow cash at no cost to buy shares, and you can also sell shares short and

(*Continued*)

(*Continued*)

pay no stock borrowing costs. In other words, all of the costs in the simulation exercise are implicit.

- You are one of five dealers: you plus the four machine dealers who are your competitors. You can trade with these dealers (interdealer trading), or you can sell shares to the public when customer buy orders arrive, and buy shares from the public when customer sell orders arrive.
- The simulation will run for one trading day.
- Your primary decision is how to set your quotes given the flow of orders that you are experiencing, and in light of your inventory position. Your may also trigger trades with the other four dealers.
- Your performance will be assessed according to two alternative metrics: (a) your closing profit and loss (P&L) and (b) a risk measure. The risk measure is the average absolute size of the inventory positions that you have carried over the course of the trading day.
- Make every effort to bring your share position back to zero by the end of each trading day so that you can avoid taking overnight risk.

For each simulation run, set the expected annual rate of return equal to 10 percent. At the start of each simulation run, set two parameters: the average time between order arrivals measured in minutes [E(ORDERS)], and the standard deviation of annual returns [SD(RETURNS)]. Run the simulation for each of the following two market settings:

| Market | E(ORDERS) | SD(RETURNS) |
|---|---|---|
| A | 3.5 minutes | 10 percent |
| B | 6.5 minutes | 5 percent |

For each market, run the simulation enough times to get a sense of the distribution of returns that you might realize from that market. Then respond to the following:

1. Which of the two market settings did you prefer to be making a market in, A or B? Be sure to explain the reason for your selection.
2. For the set of simulations that you ran, what was your largest inventory position in absolute value? How did you respond as this position was being reached? Do you think that your response kept you from realizing an even larger imbalance?

3. Contrast your two performance measures for the two different markets, A and B. For which market were your performance measures better?
4. Contrast your P&L measure over successive simulation runs. Do you feel that you did a bit better with experience? Explain why or why not.
5. Contrast your risk measure over successive simulation runs. Do you feel that you did a bit better with experience? Explain why or why not.
6. Were your realized transactions volume and market share related to the size of your inventory fluctuations?
7. For which measure, P&L or risk, did your performance improve the most? Can you explain why?
8. On the basis of your simulation experience, did you sense that accepting larger inventory swings was costly to you as a dealer? Do you understand more fully now why a dealer firm's expected average cost can be higher when it allows its inventory to fluctuate more widely? (In asking this one, we have Exhibit 5.14 in mind).
9. How easy was it for you to control your inventory positions? Did you succeed in bringing your inventory positions back down to the neighborhood of zero by the end of each simulation run?
10. Do you feel that your P&L results were attributed largely to the bid-ask spread, or was some other factor involved?

# CHAPTER 6

# Sources and Nature of Competition

Price has thus far played the key role in our analyses of household, firm, and industry behavior. Price is a critical economic signal that households and firms respond to in making their consumption and production decisions. In the previous chapters we have, with reference to price, established the demand curves of individual households, the cost curves and supply curves of individual firms, and the industry aggregate demand and supply curves.

We have not yet, however, explicitly addressed the issue of competition per se. That is the subject of the current chapter. As we will show you, the optimal price and output decision for a firm depends on the organizational structure of the industry that the firm is operating in. In the first four sections of the chapter we look at four basic structures:

1. Perfect competition
2. Monopoly
3. Imperfect competition: monopolistic competition
4. Imperfect competition: oligopoly

Our analysis of these competitive structures calls particular attention to concepts such as economic profits, economic rents, and deadweight loss; to market realities such as barriers to entry, product differentiation, and administered prices; and to strategic actions such as price discrimination and nonprice competition.

In giving more emphasis to nonprice competition, we again focus on one micro market in particular—the equity market. After providing additional background information on market maker operations and a monopoly dealer pricing model, the chapter focuses on the importance of both price and non-price competition in an (imperfectly) competitive dealer market.

## PERFECT COMPETITION

Much of our discussion in the book has been in the context of a competitive market. We will now call this environment by its more formal name, *perfect competition*. The term is a bit strange. As we have noted, two stringent assumptions characterize perfect competition: (1) firms in the industry produce a product that is fungible (as we noted in the previous chapter, the red winter wheat grown by one farmer, for example, is identical to the red winter wheat grown by any other farmer), and (2) all firms in the industry are sufficiently small so that no single firm has the power to affect price on the market (each firm, in other words, is atomistic).

So, if all perfectly competitive firms produce identical products and if no firm can influence price, how do firms compete with each other? They actually do not compete with each other in the perfectly competitive environment, which is why the term is a bit strange. They concentrate on their own operations, always seeking to minimize the height of their costs curves. To compound the puzzling use of words, a market environment where firms compete aggressively with each other, where firms seek brand recognition, where each firm claims to have the best product and fights fiercely for market share, that is the environment that we call *imperfect competition*. So much for terminology.

### The No-Profit, Long-Run, Perfectly Competitive Equilibrium Solution

We have already modeled the perfectly competitive industry. We have shown in Chapter 3 how price is set in the marketplace by the intersection of the market's demand curve for $X$ and the market's supply curve of $X$. We have shown in Chapter 5 how each firm in the industry maximizes its profits by producing at a rate that equates marginal cost with price. We have extolled the virtue of the $MC = P$ outcome because price reflects marginal utility, and it is excellent to harmonize marginal cost with marginal utility. One might even suspect that the word "perfect" has been included in the term "perfect competition" because the price and output equilibrium achieved by the perfectly competitive industry is perfectly desirable.

There is something further to say about the long-run equilibrium achieved by the perfectly competitive industry. Not only is marginal revenue equal to marginal cost, but competition drives profits to zero. Economists deem this desirable from a public policy perspective. Competition achieves this because the existence of profits attracts new firms into the industry (or the realization of losses results in an exodus). We can see how this works with reference to Exhibit 6.1.

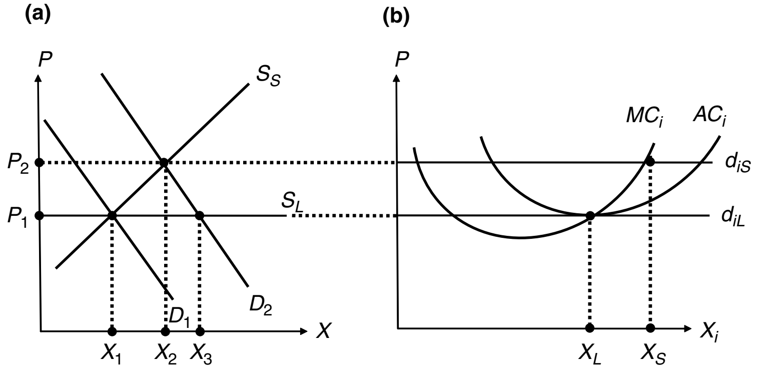


**EXHIBIT 6.1** Short-Run and Long-Run Supply Curves of a Competitive Industry [6.1 (a)], and Marginal and Average Cost Curves for Each Firm in the Industry [6.1(b)]

As we turn to Exhibit 6.1, assume for now that all actual and potential firms in the competitive industry are identical in that they have identical cost curves. The average and marginal cost curves for the representative ($i^{th}$) firm in the industry are shown in Exhibit 6.1(b). Two industry supply curves are shown in Exhibit 6.1(a). The horizontal supply curve labeled $S_L$ in Exhibit 6.1(a) is the long-run supply curve for the industry. The long-run supply curve is infinitely elastic because we have assumed that all actual and potential firms are identical, and the industry expands its output in the long run by new firms entering the industry, not by existing firms producing more.

The supply curve labeled $S_S$ in 6.1(a) is the short-run supply curve for the industry. Suppose that, in response to a demand shift, price jumps from $P_1$ to $P_2$. New firms do not have time to enter the industry in the short run. Consequently, the industry expands its output in the short run by the fixed number of firms each producing a larger output.

Two demand curves are shown in 6.1(a). The curve labeled $D_1$ represents an initial location of the industry's demand curve, and the curve labeled $D_2$ shows the curve's new location following a demand shift. With reference to the two demand curves and the two supply curves, we depict three equilibrium solutions:

1. **The initial long-run solution given by the intersection of $D_1$ and $S_L$:** Price is $P_1$ and industry output is $X_1$. Exhibit 6.1(b) shows that when price is $P_1$ and the industry is in long-run equilibrium, the $i^{th}$ firm produces $X_L$ units of $X$. The number of firms actually operating in the industry must therefore be $N_1 = X_1/X_L$.
2. **The short-run solution following a demand shift given by the intersection of $D_2$ and $S_S$:** Price is $P_2$ and industry output is $X_2$. Exhibit 6.1(b)

shows that when price is $P_2$ and the industry is in short-run equilibrium, the $i^{th}$ firm produces $X_S$ units of $X$. Because $X_S > X_L$, we have an industry output of $X_2 > X_1$ with the number of firms in the industry remaining fixed at $N_1$.

3. **The new long-run solution given by the intersection of $D_2$ and $S_L$:** Price is $P_1$ and industry output is $X_3$. Again, Exhibit 6.1(b) shows that when price is $P_1$ and the industry is in long-run equilibrium, the $i^{th}$ firm produces $X_L$ units of $X$. The number of firms in the industry is now $N_2 = X_3/X_L > N_1$.

Let's step back and see what we have. Initially the industry was in long-run equilibrium given by the point where the demand curve $D_1$ intersects the long-run supply curve $S_L$. Then demand shifts to $D_2$, and the short-run solution is given by the intersection of $S_S$ and $D_2$. In the short run, price rises to $P_2 > P_1$ and, in response, the existing firms expand their outputs from $X_L$ to $X_S$. In the short run, the firms that were already in the industry receive a profit because [look at Exhibit 6.1(b) closely], with a short-run output rate of $X_S$, $P_2$ is greater than average cost. This profit attracts additional firms into the industry; their entry expands output and lowers price. New entry continues until price is driven down to $P_1$, until profits are driven back to zero, and until the industry is again in long-run equilibrium (which now is given by the point where the industry's new demand curve $D_2$ intersects the industry's long-run supply curve, which is infinitely elastic).

## Profits and Economic Rents

What can we say about profits in this perfectly competitive industry? The profit motive leads a firm to minimize the height of its cost curves and to operate at a rate that satisfies its best profit output condition. Beyond this, profits in the competitive environment serve one role only: They are a signal that attracts new firms into the industry (or, if demand shifts to the left, negative profits force some firms to leave the industry). In long-run equilibrium, the competitive firm does not realize a profit. Of course, we could say that normal returns for services provided by the firm's owners are embedded in the firm's cost curves. As a formal matter, however, we have (in Chapter 5) assumed just two factors of production (labor and physical capital) and, at this point, we will stay within the confines of our model.

What if firms do not have identical cost curves? This is an important question. Some firms might have production advantages over other firms in the competitive industry. Agricultural land, for instance, is not of equal fertility. Some fields are easier to till than other fields. Some farms may be closer to a port, a rail head, or some other depot where crops are sold.

When firms differ in their costs, the specific firms that are in an industry are not arbitrarily selected (as they are when all firms, both actual and potential, are identical). When firms are not identical, they enter an industry according to the height of their average cost curves, with the firms with the lowest costs entering first.

For the situation that we have just described (nonidentical firms), the industry's long-run supply curve is no longer infinitely elastic; rather, it is upward-sloping. In this case, the long-run, no-profit equilibrium for the individual firm depicted in Exhibit 6.1(b) holds only for the marginal firm, the firm with the highest average cost curve of all those that are actually operating in the industry. All intramarginal firms will make excess returns; the firms with the lowest average cost curves will make the biggest excess returns. Only the excess returns realized by the marginal firm can attract new firms into the industry. Following the shift of the demand curve, new firms will enter, in an order dictated by the height of their average cost curves, until excess returns are no longer realized by the last entrant, the new marginal firm.

Notice that we have said "excess returns," not "profits." Excess returns include profits and something else. At this point economists make a distinction between *profits* and the "something else." We are talking about

$$\text{Excess returns} = [P - AC(X)]X > 0$$

where $AC(X)$ is average cost as a function of $X$ (in other words, the firm's average cost curve). If this amount can be whittled away (i.e., brought down to zero) by the entrance of new firms, it represents profits. But if it cannot be whittled away, it must be attributable to a productive advantage that is unique to the firm (for example, "better land," as we discussed before). Then the excess return is not called "profits." The "something else" is *economic rent*. In essence, the owner of the better property collects an economic rent because of the unique attributes of an asset (e.g., land) that it commands.

Let's be crystal clear about the differentiation between profits and rents. Profits are inherently short-lived, while rents for intramarginal firms are long-lived (as long, that is, as a firm remains intramarginal).[1] Positive profits attract new firms into the competitive industry while negative profits force some existing firms out of the industry. Thus profits are part of the transition mechanism that operates when an economic shock affects the relationship between price and average cost for the competitive firm. Consequently, profits endure only in the short run while the industry completes an adjustment from one equilibrium to another. Rents, on the other hand, are inherent to the individual firms. As we discuss further below, rents affect the order in which firms enter and exit an industry. And, as we have said, rents for intramarginal firms do endure when an industry is in long-run equilibrium.

---

[1] The term *intramarginal firm* refers to firms that are currently operating in the industry, except for one, the marginal firm. The term *marginal firm* applies to the last firm to enter the industry, which would also be the first firm to leave if excess returns were to decrease. The term *extramarginal firm* refers to firms that are not currently operating in the industry but that would enter if price were to rise above their own minimum average cost of producing $X$.

When the long-run supply curve of the competitive industry is infinitely elastic (that is, when all firms have identical costs), no firm enjoys economic rents. When the long-run supply curve is upward-sloping (that is, when firms are not identical in terms of costs), the intramarginal firms realize excess returns in the long run. As we have said, these excess returns are called *economic rents*. An interesting thing about economic rents is that, if they can be identified by an eagle-eyed government accountant, they could be taxed away without changing the firm's output decision. This is because the goal of the firm is to maximize its economic rents and, whatever the output might be that maximizes its rents in the absence of a tax, continues to maximize its rents in the presence of a tax (because all values of this net return, as a function of $X$, are reduced proportionately).

In other words, an economic rent tax would simply level the playing field from a cost point of view for all of the firms in a competitive industry. Such a tax would be the least obtrusive levy that a government could impose. Unfortunately, economic rents are not easy to identify, and what might sound good in theory is not necessarily implementable in practice. On the other hand, the relative cost advantages of different firms can (at least to some extent) be known to other market participants, and firms that enjoy lower costs will command higher valuations. If we allow for a secondary market that enables the purchase and sale of firms, the value of any firm's economic rent would be capitalized in the price of the firm that would be paid by a new owner. Real estate provides an example: whether we are talking about farms, industrial plants, or residential homes, good location matters and it is paid for.

And so, whether or not the firms in a competitive industry are identical, profits do not exist in the long run. Profits are either eliminated by the entry of new firms, or they are not called profits (they are called economic rents that could be taxed away without altering any individual firm's output decision). You might wonder about this. You might ask if profits could serve any economic function other than to attract new firms into an industry whenever price is greater than average cost for the marginal firm. Do profits, for instance, represent a salary paid to the owner of a firm in a competitive industry?

The answer to this one is no, they do not. Any resource that is required to run a firm is a factor of production, and any salary paid for that factor should be included in the firm's cost curve. Our measure of profits is not included in the cost curve. Our measure is the difference between total revenue and *total cost* (which is the same as output times the difference between price and average cost).

But profits do serve an economic function, and it is twofold. First, profits are a compensation for risk taking. Second, profits are a reward for innovating, for pioneering with regard to product development and cost-saving production techniques. But neither risk nor innovation is included in our deterministic (nonstochastic), comparative static (not dynamic) analysis. Without risk or innovation in the model, there is no economic function left for profits to play other than to attract new entrants into an industry and, in so doing, to compete any profits away, or to force existing firms out of the industry if they are operating at a loss.

### Deadweight Loss

We have considered a competitive market's economic efficiency in terms of its ability to compete excess returns away. The economic efficiency of a perfectly competitive market can also be assessed in terms of the elimination of *deadweight loss* that occurs when a free market equilibrium is achieved. In this subsection, we define deadweight loss and show its occurrence. We do so with reference to Exhibit 6.2.

Suppose that sales of product $X$ are taxed, and that the tax is a flat-dollar amount per unit sold. The effect of this tax is shown in Exhibit 6.2 (which largely replicates the sales-tax diagram in Exhibit 3.10 and the dealer pricing model in Exhibit 5.12). Without the tax, the intersection of the supply curve $S_1$ and the demand curve $D$ establishes $P_1$ as the equilibrium price and $X_1$ as the equilibrium quantity of $X$. A sales tax of $T$ per unit results in a parallel shift of the supply curve up and to the left, by the amount $T$, to the curve labeled $S_2$. (Alternatively, we could have shifted the demand curve down and to the left by the amount $T$ and held the supply curve constant.) The choice is arbitrary. In fact, neither supply nor demand has actually shifted. Rather, from the perspective of buyers, it is "as if" the supply curve has shifted and, from the perspective of sellers, it is "as if" the demand curve has shifted. (Look again at the discussion in Chapter 3).

With the sales tax, the intersection of $S_2$ and $D$ establishes $X_2 < X_1$ as the equilibrium quantity, $P_{2B} > P_1$ as the price that buyers pay, and $P_{2S} < P_1$ as the price that sellers receive. Buyers paying a higher price and sellers

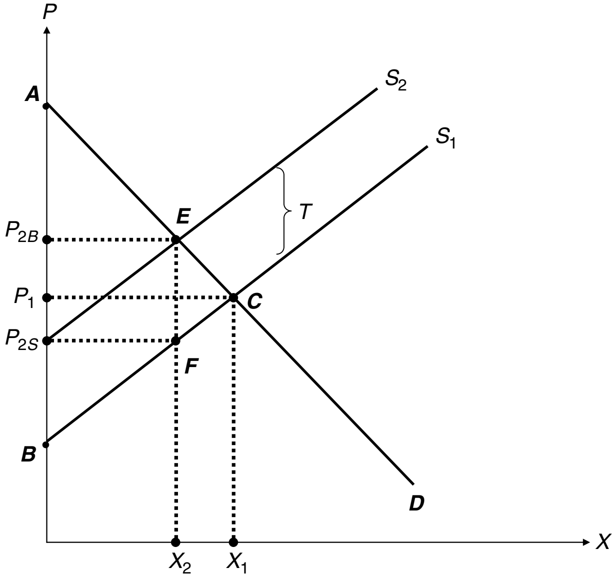


**EXHIBIT 6.2** Deadweight Loss from the Imposition of a Sales Tax

receiving a lower price explains output decreasing to $X_2 < X_1$. Can we quantify the cost to buyers and sellers and compare it to the revenue received from the tax? We can, using a dollar measure of the gains from trading.

The dollar measure of the gains from trading, for buyers, is *consumer surplus*. We identified consumer surplus in Chapter 4 and used the concept to analyze the trading decision of an individual placing an order to trade shares of a stock. From the seller's point of view, *producer surplus* is an analogous concept. Here are the two concepts, spelled out side-by-side:

**Consumer surplus:** When an intramarginal unit of $X$ is worth more to buyers than the price paid (for instance, $P_1$ in Exhibit 6.2), the buyers benefit by a monetary amount equal to the height of the demand curve for that intramarginal unit minus the price actually paid which we are taking to be $P_1$. As shown in Exhibit 6.2, applying this measure to all intramarginal units, consumer surplus without the tax is the area $ACP_1$ when price is $P_1$.

**Producer surplus:** When an intramarginal unit of $X$ can be produced by sellers at a lower cost than the price received ($P_1$ in Exhibit 6.2), the sellers benefit by a monetary amount equal to the price actually received (which we are taking to be $P_1$) minus the height of the supply

> curve for that intramarginal unit. As shown in Exhibit 6.2, applying this measure to all intramarginal units, producer surplus without the tax is the area $P_1CB$ when price is $P_1$.

We have just identified consumer and producer surplus for the no-tax equilibrium. Adding the two areas together, the gains from trade for both buyers and sellers is the larger area, $ACB$ in Exhibit 6.2. With a tax, consumer surplus is reduced to the area $AEP_{2B}$, producer surplus is reduced to the area $P_{2S}FB$, and the tax revenue is $P_{2B}EFP_{2S}$. Thus the monetary gains from trade (with the taxing authority included as a beneficiary) is the area $AEFB$. Contrasting the gains from trade without a tax with the gains from trade with a tax we have:

$$\begin{aligned}\text{Gains from trade without tax} - \text{Gains from trade with tax} &= ACB - AEFB \\ &= ECF\end{aligned}$$

The area $ECF$ is a reduction in the gains from trading that has no offset. It is purely and simply a loss. Accordingly, the area $ECF$ is called a *deadweight loss*. In a frictionless trading environment, free market equilibrium for a competitive industry involves no deadweight loss. This is another way to appreciate the efficiency of the frictionless, free-market solution in a perfectly competitive environment.

## MONOPOLY

At one end of the competitive spectrum we have perfect competition. At the other end we have monopoly. In the middle is a rather odd collection of situations that are collectively referred to as imperfect competition. Our purest, most straightforward formulations are for perfect competition and monopoly. In this section, we focus on the latter.

### Definition of Industry

We continue to call our generic consumption good $X$. When $X$ is produced by just one firm, the firm is a monopolist and the industry is monopolized. With just a single firm in the industry, the demand curve facing the firm is the same as the demand curve facing the industry. Because the monopoly firm has no competition from within its industry, it has the power to set price (as well as output). But do not think that the monopoly firm does not face competition. It does have to compete. Its product, $X$, competes with all other items in household consumption baskets. The key consideration is that no other good is a close enough substitute for $X$ for it to be considered in the same industry as $X$.

This is a tricky one, however: *an industry* may not be identifiable without ambiguity. Perhaps $X$ is a Microsoft product, Microsoft Office. No one else produces the Word-PowerPoint-Excel package. Is Microsoft a monopolist? Is the product (the industry) Microsoft Office, or is it a broader set of PC software? What other software items along with Microsoft Office should be included in the definition of the industry?

A firm may also have monopoly power in a local market but not in a national market. A small and relatively isolated town, for instance, may have only one gas station, one supermarket, and one dry cleaner because the market is not big enough to support more than one of each. Accordingly, these firms have some monopoly power. Consequently, they can charge higher prices.

Returning to our generic product, $X$, the question is whether or not there are close enough substitutes for $X$. If a close substitute for $X$ (let's call it $Y$) exists, and $P_X$ changes, $P_Y$ is likely to respond. But if $P_Y$ does change, how will the response affect the demand for $X$? If the producer of $X$ takes the responsive change of $P_Y$ into account, the monopoly model loses its sharpness. If the interaction between $P_X$ and $P_Y$ is strong enough, we should expand our definition of the product and consider the industry imperfectly competitive rather than monopolized.

In Chapter 3 we identified two goods as being substitutes or complements according to their cross elasticity of demand, $\eta_{X \cdot P_y}$ (look again at Chapter 3's equation 7). And so, to assess $Y$ as a substitute for $X$, take the cross elasticity of the demand for $Y$ with respect to the price of $X$. Because $Y$ is a substitute for $X$, $\eta_{Y \cdot P_X} = \frac{dY}{Y} \div \frac{dP_X}{P_X} = \left(\frac{dY}{dP_X}\right)\left(\frac{P_X}{Y}\right) > 0$. If $\eta_{Y \cdot P_x}$ is sufficiently small, substitution is weak and the industry producing $X$ can be considered monopolized. If $\eta_{Y \cdot P_x}$ is sufficiently large, substitution is strong and the industry should be defined more broadly so as to include $Y$. We used the word "sufficiently." There is no clear break point for the cross elasticity term. The final assessment calls for the use of judgment.

## The Monopoly Firm's Best Profit Price and Output

As is the goal of a competitive firm (or any other firm for that matter), the goal of the monopolist is to maximize profits. The monopoly firm's cost curves are derived from its production function just as they are for a competitive firm. The monopoly firm's best profit output is also similarly determined: produce at the rate that equates marginal revenue with marginal cost. But, unlike the competitive firm, marginal revenue does not equal average revenue (price) for the monopolist. This is because the monopoly firm faces a downward sloping demand curve.

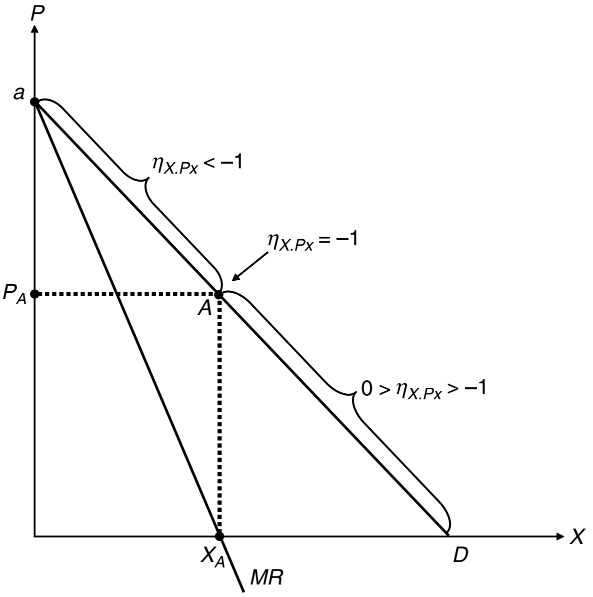


**EXHIBIT 6.3** Demand and Marginal Revenue Curves of a Monopoly Firm

In Chapter 5, Exhibit 5.11, we present a monopoly diagram, although it is not labeled as such. The demand curve in that figure is, by intention, quite elastic. Our purpose in Chapter 5 was to explain that a firm has a supply curve only when the demand curve facing the firm is *infinitely* elastic, as it is in a perfectly competitive industry. Now we examine the downward sloping demand curve in its own right. We do so with reference to Exhibit 6.3.

The figure in Exhibit 6.3 presents a demand curve and its associated marginal revenue curve. When demand is linear (as it is in the exhibit), marginal revenue is also linear, and it decreases with $X$ at twice the rate as does demand. Here is how you can explain it:

1. Write the demand curve for $X$ as $P_X = a - bX$.
2. Multiply both sides by $X$ to get total revenue: $TR = P_X X = aX - bX^2$.
3. Marginal revenue, the derivative of total revenue with respect to $X$ (if you remember your calculus) is $MR = \frac{d(P_X X)}{dX} = a - 2bX$.
4. From the demand curve we have $\frac{dP_X}{dX} = -b$. From the marginal revenue curve we have $\frac{dMR}{dX} = -2b$. This completes the demonstration.

In Chapter 5, we obtained the relationship between $MR$, $P$, and $\eta_{X \cdot P}$. It is $MR = P\left(1 + \frac{1}{\eta_{X \cdot P}}\right)$ (look again at equation 5.10). From this relationship, we have

- $MR > 0$ for $\eta_{X \cdot P} < -1$ (elastic demand)
- $MR = 0$ for $\eta_{X \cdot P} = -1$ (unitary elasticity)
- $MR < 0$ for $0 > \eta_{X \cdot P} > -1$ (inelastic demand)

This shows that marginal revenue is positive for the monopolist only in the region where demand is elastic. This is shown in Exhibit 6.3.

This is a good time to call your attention to how elasticity changes as we move down and to the right along the negatively inclined, linear demand curve. As we walk you through this, keep the following equation for elasticity in mind: $\eta_{X \cdot P_X} = \left(\frac{dX}{dP_X}\right)\left(\frac{P_X}{X}\right)$.

1. Because the demand curve in Exhibit 6.3 is linear, $dX/dP_X$ is constant as we move down the curve.
2. Because $MR$ decreases with $X$ at twice the rate that $P$ decreases with $X$, the horizontal intercept for the $MR$ curve is half of the horizontal intercept for the demand curve, $D$. Likewise, the falling $MR$ curve in Exhibit 6.3 bisects the straight line from $P_A$ to point $A$.
3. As we move down the curve, $P_X$ falls and $X$ increases. Therefore, the ratio $P_X/X$ falls and so too does elasticity.
4. As we move up the curve and approach the vertical intercept (the point labeled "*a*" in the diagram), $X$ goes to zero and the elasticity of demand goes to (negative) infinity.
5. As we descend from the intercept (labeled "*a*") to the point where $MR = 0$, which is point $A$ (with coordinates $P_A$ and $X_A$), elasticity falls to (minus) unity; see equation (5.10).
6. Further down the demand curve, past point $A$, demand is inelastic and marginal revenue is negative.

The monopolist, in maximizing profits, produces at an output rate where $MR = MC$, and $MC$ is always positive. Therefore, in equilibrium, $MR$ must be positive for the profit-maximizing monopoly firm. Accordingly, a monopoly always operates in the elastic part of its demand curve. This is not the case for the competitive industry where equilibrium is determined by the point where demand crosses supply, wherever that crossing point might be.

The monopolist operating in the elastic part of the demand curve can be understood intuitively. If the monopoly firm is producing at a rate that locates it at any point in the inelastic part of the demand curve, it can raise price without having to reduce output proportionately (which is what inelasticity means: the percentage change in quantity is less than the percentage change in price). Accordingly, total revenue (price times quantity) goes up and, with output being lower, total costs must be less. Therefore, with revenue higher and costs lower, profits are higher. Consequently, profits may be positive, but they cannot be at a maximum at any point where demand is price inelastic.

The fact that the firm operates in the elastic part of the demand curve has an interesting application to the pricing of books, and to the relationship between authors and their publishers.[2] Author royalties are a percent of the total revenue received by the publisher and, accordingly, authors want total revenue maximized. How is this done? By pricing a book so that its sale price is at the value where marginal revenue is zero. With reference to Exhibit 6.3 (assuming that $X$ is a book), the best price from the author's point of view is $P_A$. As we see in Exhibit 6.3 (and also know from equation 5.12), demand elasticity equals one when marginal revenue equals zero and total revenue is maximized. But the publisher sets the book's price higher so as to equate marginal revenue with a marginal cost that is positive. In so doing, the publisher wants to be in the elastic part of the demand curve. Consequently, authors and publishers do not see eye to eye on this one. Lower prices are better for authors, and higher prices are better for publishers. We trust that our publisher will forgive us for pointing this out.

### The Price-Discriminating Monopolist

The prices that are set in monopoly, monopolistic competition, and oligopoly markets are *administered prices*. In setting an administered price, a firm may be able to divide its customers into groups and charge the groups separately according to their differential sensitivities to price. For the firm to maximize its profits, customers who have good substitutes available should be charged lower prices, and customers for whom good substitutes do not exist should be charged higher prices. Segmenting customers into groups and charging them according to their elasticities of demand is called *price discrimination*.

Profitable price discrimination requires three things:

1. It should be legal. Some applications of this pricing strategy may violate the 1936 Robinson-Patman Act.

---

[2] At this point we can raise the following question. Is the publisher of a book a monopolist with respect to that book? The answer is "yes" if one simply considers that no other publisher has the right to sell the book to bookstores or other distributors. However, what is the answer when you recognize that different books can be substitutes for each other? This brings us back to a thought expressed earlier in this section: The definition of a product and the industry that produces it is not unambiguous.

2. It must be possible for the firm to segment its customers into two or more groups that cannot trade with each other, either directly or via the intermediation of an arbitrageur. Arbitrage trading would cause any price differential to break down.
3. The differential prices charged to the different groups must be attributable not to cost differentials (for that would not be price discrimination), but to the customers in the different groups having different demand elasticities. We will show you that, when costs are the same, it is optimal to charge the different groups different prices only if their elasticities differ.

Two examples of price discrimination are railroads and movie theaters that charge different prices for riding the rails, or for watching movies, at different times of the day. Commuters who have to get to their offices in the morning do not find noon-hour trains a good substitute for earlier trains, and their demand for the earlier hour trains is, therefore, relatively price inelastic. People with 9 to 5 jobs do not have the option of seeing daytime movies during the workweek, and thus their demand for evening time shows is relatively inelastic (because daytime shows are not a good substitute). But retired people can travel and see movies more readily at various times during the day, and their demand for these daytime activities can be relatively price elastic. So, the optimal strategy for a railroad or a movie theater is to lower the price of the daytime event (for which demand is relatively elastic) and to raise it for the rush hour or evening event (for which demand is relatively inelastic).

Can these two markets be successfully segmented? They certainly can be. Tickets are time-stamped, and train rides and movies are services that, unlike physical goods, cannot be bought in one place, transported through space or time, and then sold somewhere else.

If a market can be segmented into two different groups, one with a relatively inelastic demand and the other with a relatively elastic demand, the monopoly firm can increase its total revenue for the same unit sales volume by taking a unit away from the relatively inelastic segment, where its marginal revenue is relatively low, and redirecting it to the more elastic segment, where its marginal revenue is relatively high. This redirecting of sales should continue until the revenues the firm receives from its two market segments are equal *on the margin*.

But for revenues to be equal on the margin when the elasticities of demand are different, prices in the different market segments must be different. We can see this for two segments by using an equation that we have put to good use before in Chapter 5, equation 5.10: $MR = P\left(1 + \frac{1}{\eta_{X \cdot P}}\right)$. Using the letters $A$ and $B$ to identify the two markets and suppressing the $X.P$ subscript on $\eta$, we have

$$MR_A = P_A\left(1+\frac{1}{\eta_A}\right) = MR_B = P_B\left(1+\frac{1}{\eta_B}\right) \qquad (6.1)$$

For equation 6.1 to hold, if $0 > \eta_A > \eta_B$, we must have $P_B < P_A$ (because $\eta_B$ is a larger negative number, its reciprocal is a smaller negative number, and the value of the parenthesized expression is therefore larger, which means that the price which the parenthesized expression is multiplied by must be lower to preserve equality). This completes the algebraic demonstration.

Perhaps you would like a graphical demonstration of this result. Turn to Exhibit 6.4. Exhibit 6.4(a) shows the demand curve (labeled $D_1$) for one segment of the market, 6.4(b) shows the demand curve (labeled $D_2$) for the second segment of the market, and 6.4(c) shows the aggregated demand curve, $D_1 + D_2$. To preserve linearity of the aggregated demand curve, we have located the two disaggregated demand curves with the same price intercept. Because of this, and given that $D_1$ is falling faster than $D_2$, the demand curve in 6.4(a) is less elastic than the demand curve in 6.4(b) (that is, at any value of $X$, we have $0 > \eta_1 > \eta_2$).

The monopolist's marginal cost curve is shown along with the aggregated demand curve in 6.4(c). If the monopolist does not price discriminate, it sells $X^*$ units at a price of $P^*$. If, alternatively, the firm divides its market into two segments, it can set marginal revenue in each segment equal to the

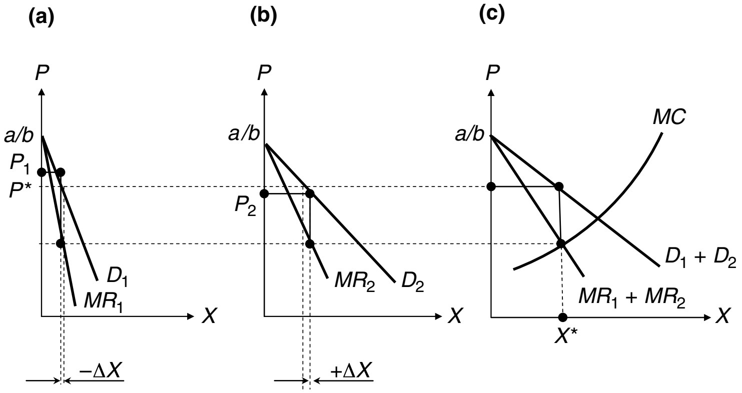


**EXHIBIT 6.4** Demand and Marginal Revenue Curves of a Price-Discriminating Monopolist for a Relatively Price-Inelastic [6.4(a)] and a Relatively Price-Elastic [6.4(b)] Market Segment, with Total Output Set by the Aggregate Marginal Revenue and Marginal Cost Curves [6.4(c)]

common marginal cost that is established in 6.4(c). As the firm moves units from the market segment where marginal revenue is lower to the market segment where marginal revenue is higher, its profits increase. Accordingly, the firm sells $\Delta X$ fewer units in the segment where it raises its price to $P_1$ [as shown in 6.4(a)], and it sells $\Delta X$ more units in the segment where it lowers its price to $P_2$ [as shown in 6.4(b)]. By implementing this pricing strategy, the firm loses less revenue in segment 1 (or perhaps it even gains revenue) than it gains in segment 2 (because demand is less elastic in segment 1 than in segment 2). This is what enables the price discriminating monopoly to realize a greater profit.

### What Is the Problem with Monopolies?

The problem with monopolies, in a nutshell, is that industry output is smaller and price is higher in a monopolized market than in a competitive market. In a competitive industry, as we see in Chapter 5, firms maximize their profits by producing amounts that set their marginal costs equal to price (because, as we have established, with infinitely elastic demand on the firm level, price equals marginal revenue which in turn equals marginal cost). Also, as we have just noted, marginal cost can equal price for a competitive firm at an industry output level where the demand curve facing the industry is inelastic. In contrast, a monopoly only operates in the elastic part of its demand curve.

With this in mind, take a competitive industry, with its industry demand curve and cost curves, and monopolize it. Do it so that there is only one change—a change in ownership. That is, *ceteris paribus*, let one big firm buy all $N$ little firms in the industry and then keep them operating as if they were separate plants. Or assume that an industry organization like the Organization of Petroleum Exporting Countries (OPEC) sets output levels for all of its member firms. This *ceteris paribus* change of ownership or control moves the output decision from the firm level to the industry level. What happens? Output is reduced and price is increased. The new equilibrium price is not only higher than it was before; price is also higher than marginal cost. This is a major reason why economists inherently prefer competition over monopoly. "Keep the markets competitive" is the mantra for many economists and government regulators alike.

Notice, however, that this monopoly, perfect competition contrast has been made within the context of a static economic model. In a dynamic environment, monopoly power can have some desirable features. For one thing, competition aimed at achieving a monopoly position can itself lead to dynamic product innovation and development. We return to this thought later in the chapter. Also, a monopoly firm can better exploit economies of scale if the economies extend over a large enough region of output. We return to this thought right after the next subsection.

### Barriers to Entry

All firms want control over the price of their product. They all wish to realize the profits associated with having monopoly power over price. What enables a firm to achieve this hallowed position? The answer is three words long: *barriers to entry*.

Various barriers can keep competition from entering an industry. A firm that has invented a new product will (for some time at least) be the sole producer of that product. Keeping formulas and production techniques well guarded secrets helps to maintain a protected position (only Coca-Cola knows the formula for producing Coca-Cola). Exclusivity may be achieved by the use of trademarks, copyright protection, and patents (the term of a patent is generally 20 years). Control over a vital resource such as a privileged geographic location may make it difficult for new competitors to enter a field. A government may extend an exclusive license to a firm, or to a franchise such as the National Football League. Economies of scale in production are a major and particularly important barrier to entry. We consider this one in its own right in the subsection that follows. First, we have an important clarification to make.

Competition does not have to be actual for competitive pressures to be applied to a monopoly firm. The very threat of entry may be enough to keep a monopoly firm's price reasonable. *Contestability*, or *contestable markets*, is the term that economists use when a single firm in an industry has to keep its price relatively low to hold potential competition at bay.

### Natural Monopoly

One of the most effective nongovernmental barriers to entry is economies of scale in production. For a firm that enjoys economies of scale, average costs fall as its output expands. This means that, to exploit the efficiencies offered by scale economies, the firm must increase its output. If the economies of scale extend to output levels that are large relative to the aggregate industry, then one firm, in exploiting the scale economies, will come to dominate the industry. It will do so for a simple reason: the economies enable one firm to produce output for the entire industry at a lower cost than any two or more other firms can together achieve. This being the case, other firms operating at lower output levels are less efficient and, being less efficient, they cannot compete. They drop out and one firm survives as a monopolist.

As we discuss in Chapter 5, economies of scale means that the average cost of producing $X$ decreases as the output of $X$ is increased. Two other concepts are related to economies of scale: *economies of scope* and *network externalities*.

1. **Economies of scope:** While economies of scale apply to the production of one product (for instance, our generic product, $X$), economies of scope is the applicable term when a number of different products or activities are involved. The ability to use a sales force more effectively when selling an array of products is an economy of scope. Or a firm that has a variety of different product lines can give its graphic designers more efficient work schedules (i.e., keep them better occupied), and this lowers the cost of producing each product. Another example is the risk reduction that is achieved by offering a range of products (just as diversifying the assets in any portfolio reduces risk).
2. **Network externalities:** A network exists when one user of a good or service bestows benefits that can be either positive or negative on other users. Network effects are generally positive, and the larger a network is, the greater is the value that it offers its members. We call it a network *externality* because an individual, when joining a network, does not take into account the effect that his or her entry has on the value of the network to others who are already in it.

Telephones and fax machines are good examples of a network. Each individual's telephone or fax is more useful the more people there are who have a telephone or fax, because there are more people to whom calls can be made, to whom faxes can be sent, and from whom calls and faxes can be received. A larger number of people using the same software also generates network value; we all benefit from the ability to pass files back and forth to one another in a standardized language and format.

Similarly, an equity market is a network. The network comprises the participants who come together to trade. Each participant is a node in the network. As the number of nodes increases, a market becomes more liquid, and its performance characteristics improve. An adage in the securities industry reflects the positive force that a network exerts: "order flow attracts order flow." The equity market network does this because bigger markets generally offer more liquidity, tighter spreads, and better price discovery. But each trader, individually, does not take account of the extent to which his or her own orders are adding to the quality of the network. For each participant individually, the benefits bestowed on others are an externality. For this reason, this economy to size is called a *network externality*.

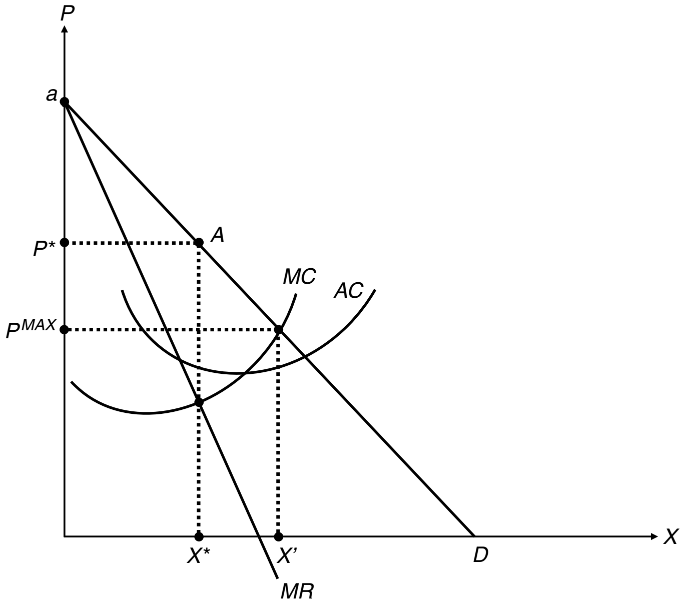


**EXHIBIT 6.5** Socially Desirable Price and Quantity Equilibrium for a Monopoly Firm Attained by the Imposition of a Maximum Price ($P^{MAX}$)

Now let's get back to the natural monopoly. We have seen that an industry may be monopolized because of the efficiencies realized from large-scale production. The power of order flow to attract more order flow in equity trading translates into an equity marketplace possessing natural monopoly properties. Indeed, there is a strong tendency for equity markets to consolidate, even while technology developments and regulatory pressures may lead to order flow fragmenting over multiple market centers. We discuss this further in Chapter 8. More conventionally, economies of scale are recognized as a technological property of a production function, as we discuss in Chapter 5 (go back to Exhibit 5.2a and focus on region *A*). We will continue our analysis in this context, with reference to an average cost curve that, due to its underlying production function, is decreasing over a large range of output. Such a curve is shown in Exhibit 6.5.

The downward-sloping linear demand curve labeled *D* in Exhibit 6.5 is the market's aggregate demand curve for *X*, and the linear curve labeled *MR* is the marginal revenue curve that is associated with that demand curve. The firm's U-shaped average and marginal cost curves are labeled *AC* and *MC*, respectively. The important thing to note is that the average cost curve is falling over a large range of output.

For the monopoly firm, as for any other firm, the best profit output is determined by the point where marginal revenue equals marginal cost (that is, where the *MR* curve intersects the *MC* curve). This equality is achieved at the output rate $X^*$. The demand curve shows that, at $X^*$, the price the monopolist will charge is $P^*$. Note that, because the average cost curve is falling as output increases (in the diagram, *AC* is even falling past the quantity $X^*$), dividing the output $X^*$ between two or more firms that face similar *AC* curves would increase average production costs for each of them. That is why the diagram applies to a *natural monopoly*. From a production point of view, having one monopoly firm is the most efficient.

But, with regard to the amount of output that it brings to market, the monopoly firm is not efficient from a public policy perspective. In the competitive industry, output, as we have seen, would be increased to the point where the marginal cost curve intersects the demand curve, and the desired equality between marginal cost and price is achieved. In Exhibit 6.5, the intersection of the curves labeled *MC* and *D* establishes $P^{MAX}$ and $X'$ as the socially desired price and quantity equilibrium.

We have labeled the socially desired price $P^{MAX}$ for the following reason. If a government regulatory agency were to know the firm's cost curves and the industry demand curve, it could achieve the desired outcome in a very simple way. It could use its authority to impose a maximum price, $P^{MAX}$. Imposition of this maximum price effectively changes the demand curve for the natural monopoly, making the demand curve the horizontal, infinitely elastic line with a vertical intercept at $P^{MAX}$ and extending out to the quantity $X'$. Then at $X'$, the downward sloping demand curve, *D*, again applies. Thus the best profit output for the regulated monopoly firm would be $X'$, where $MC = P^{MAX}$.

One could say that government, in imposing the price ceiling at $P^{MAX}$, is in effect playing the role of surrogate competition. How effectively a government agency should or could exercise this role, whether or not any preferable alternatives might exist, and how this would all play out in a dynamic environment characterized by technological change, is open to much discussion and debate.

## IMPERFECT COMPETITION: MONOPOLISTIC COMPETITION

Imperfect competition is the term we apply to markets that lie somewhere on the scale between perfect competition on one end, and monopoly on the other. Two broad classifications fit under the term *imperfect competition*:

monopolistic competition and oligopoly. We deal with these, respectively, in this section and in the section that follows.

The term *monopolistic competition* is an oxymoron: a monopolist has no direct competitors, and an industry that comprises competitors is not a monopoly. On the other hand, there is justification for the term. The "competition" part is clear. Firms in an industry characterized by monopolistic competition do indeed compete with one another. The "monopoly" part is explained as follows. Each firm in the industry has something distinct about itself and, with regard to its distinctive features, the firm is a monopolist.

An industry that fits the monopolistic competition model has two salient characteristics:

1. A large enough number of firms must be in the industry, and barriers to entry must be sufficiently low, so that (a) no one firm has dominant economic power, (b) interfirm collusion is difficult, and (c) direct competitive interactions between any two or more of the individual firms are weak. The first two conditions distinguish monopolistic competition from the monopoly model, and the third condition distinguishes it from the oligopoly model (we get to the oligopoly structure in the next section).
2. The output of each firm in the industry must, to some extent, be different from the output of the other firms. This characteristic differentiates monopolistic competition from the perfectly competitive model.

Competition can be intense under monopolistic competition. A monopolistically competitive firm will strive to differentiate its product better, to advertise, and to have sales promotions, all in an attempt to establish a brand name that distinguishes it from its competition. Product quality and good customer relationships count heavily in this effort. Customers will, for instance, choose between grocery stores, or between drug stores, or between liquor stores based on the personalities of the people working in them, the discounts that they offer, the hours that they are open, and their willingness to make home deliveries and to accept returned merchandise.

Equilibrium under monopolistic competition can be better depicted for a firm in the industry than for the industry in aggregate. Because the output of one firm is not identical to the output of another firm, aggregating unit output for the industry is similar to adding apples and oranges. Because prices differ to some extent from firm to firm, a single, unambiguous price does not exist for the industry and a price index or average would have to be used. In light of these difficulties, we go directly to the firm level of analysis. Let's turn to Exhibit 6.6 and consider the equilibrium for an individual firm that produces its own slightly differentiated version of $X$.

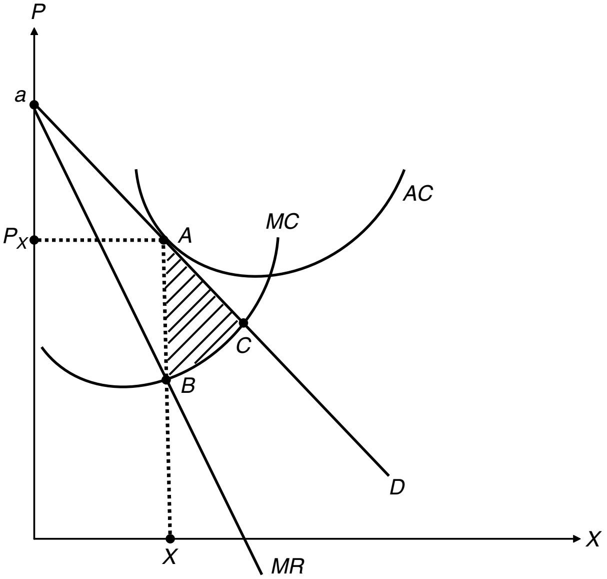


**EXHIBIT 6.6** A Downward-Sloping Demand Curve and Price, Quantity Equilibrium for a Monopolistically Competitive Firm

Exhibit 6.6 contains practically the same information as Exhibit 5.11. Although we did not say so at the time, Exhibit 5.11, in fact, is the diagram for a firm in a monopolistically competitive industry. We say in Chapter 5 that the output of one firm may be distinguishable from the output of its competitors, that a grocery, newspaper, or liquor store may compete with similar stores in its neighborhood while retaining some individuality and some independent pricing ability. We note that the demand curve facing such a firm, although relatively elastic, is not infinitely elastic. What we described in Chapter 5 is indeed a firm facing monopolistic competition.

The equilibrium solution depicted in Exhibit 5.11 is best interpreted as a short-run equilibrium. The reason for this, if you look closely at the figure, is that, given the relationship between the average cost curve and the demand curve, our representative firm is making an economic profit. Its profit results from its demand curve (*d*) cutting through its average cost curve (*AC*) so that, at the best profit output ($X^*$), price is higher than *AC*. As is the case in the perfectly competitive industry, this economic profit gives new firms an incentive to enter the industry. As they do, industry-wide output increases and the demand curve facing the representative firm in the industry shifts down and to the left.

In Exhibit 6.6, the firm's demand curve, which is labeled *D*, has shifted down and to the left until it is tangent to the firm's *AC* curve. The diagram, therefore, is for the marginal firm in the industry. If the demand curve were to shift further down, the firm would be operating at a negative profit and, in the long run at least, would leave the industry. If, in long-run equilibrium, the demand curve does not shift down enough to be tangent to the *AC* curve, the firm will continue to make excess returns that we would call not profits, but rent. If, in the long run, the firm is making zero profits, it has the dubious honor of being the marginal firm in the industry.

The tangency solution in Exhibit 6.6 suggests an interesting implication. Because the demand curve is downward sloping, tangency must be at a point along the *AC* curve where the average cost curve is also downward sloping. Consequently, in long-run monopolistic competition equilibrium, (1) average costs are not minimized, and (2) price is higher than marginal cost for the marginal firm. In contrast, as we have established, *AC* is minimized and *P* does equal *MC* for the marginal firm in a perfectly competitive industry [see Exhibit 6.1(b)].

*AC* not being minimized and $P > MC$ results in a deadweight loss. The loss is shown with reference to the marginal firm in Exhibit 6.6 by the shaded area demarcated by the points *A*, *B*, and *C*. This deadweight loss cannot be attributed to any price regulation, to any output constraint, or to any market friction. It is due simply to the fact that consumers like variety, that they like choice. The less important any difference is between the *X* produced by one firm and the *X* of any other firm, the more elastic the demand for each firm's *X* would be, the closer equilibrium would be to the competitive equilibrium identified by point *C*, and the smaller the deadweight loss would be. But product differentiation and consumer tastes being what they are, the demand curve *D* is downward sloping and the equilibrium is given not by point *C*, but by point *A*. In such a market, the participants accept the deadweight loss.

## IMPERFECT COMPETITION: OLIGOPOLY

The term *oligopoly*, with its Greek root "oligo," which means few, is applied to an industry that is comprised of, or at least dominated by, a small number of sellers. An industry that fits the oligopoly model has two distinguishing characteristics:

1. More than one firm is in the industry, which distinguishes an oligopoly from a monopoly. A special case of oligopoly is duopoly, which means that there are two firms in the industry. Duopoly is a

mathematically convenient structure to analyze, and we turn to it later in this section.

2. The number of firms is small enough for individual firms to have prominence in the industry. Having prominence, each takes account of its competitors' strategic responses when making its own strategic decisions. This characteristic distinguishes oligopoly from monopolistic competition. Unlike under monopolistic competition, the outputs of firms in an oligopoly may or may not be homogeneous, fungible products.

Oligopoly is a market form that exists because of either legal restrictions to entry (such as a government license being required, as is the case for taxi companies) or due to the existence of economies of scale that extend just far enough for the output of a firm to be large relative to the total market, but not large enough for one firm to dominate. For instance, let $AC^{MIN}$ be the value of average costs at the minimum point of a firm's average cost curve, and assume that, at a price equal to $AC^{MIN}$, the aggregate industry demand for $X$ is $X_1$. If the firms have identical average cost curves, the number of firms that the industry can efficiently support is $N = X_1/X_i$, where $X_i$ is the output of the $i^{th}$ firm when that firm is operating at the minimum point of its average cost curve. For $N > 1$ but still sufficiently small for the individual firms to have market power, we are dealing with oligopoly.

## Administered Prices and Multifirm Equilibrium

Firms in imperfect competition and monopoly industries administer their own prices. That is, taking all competitive and market factors into account, each firm determines the specific prices to put on the goods and services that it sells. The process itself involves costs: decisions have to be made, labels, menus, and/or signs have to be prepared and updated, and the price information has to be distributed to customers (among other ways, through advertising). All of this means that administered prices do not change with the same rapidity as prices in commodities markets, which are constantly responding to the shifting pressures of demand and supply.

The relative constancy of administered prices under oligopolistic competition in particular can also be explained in a different way, in a way that shines an interesting light on the interplay between competitive responses in this market environment. The formulation is called the *kinked demand curve model*. We set this model forth with reference to Exhibits 6.7 and 6.8.

As we have noted, firms in an oligopolistic industry can produce either homogeneous or differentiated products (in contrast to monopolistically competitive firms, which must produce differentiated products). Assume a

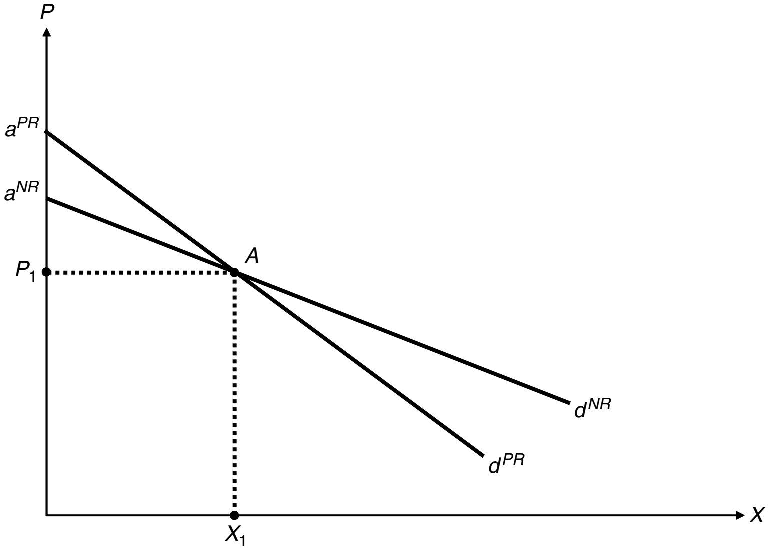


**EXHIBIT 6.7** A No-Response Demand Curve $d^{NR}$ and a Perfect Response Demand Curve $d^{PR}$ for an Oligopolistic Firm

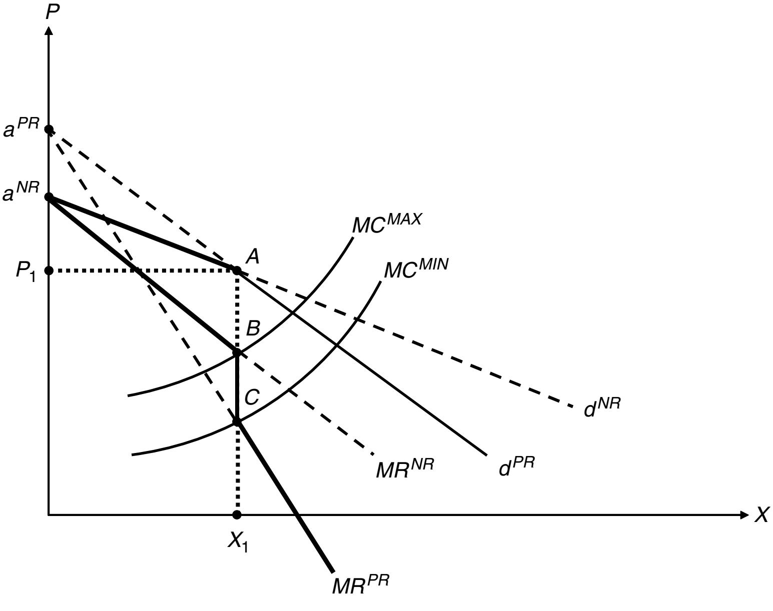


**EXHIBIT 6.8** The Kinked Demand Curve for an Oligopolistic Firm

small number of oligopolistic firms that, for the sake of simplicity, produce a homogeneous product, $X$, and consider the demand curve that is faced by any one of these firms. For starters, as shown in Exhibit 6.7, assume that an administered price has already been set at $P_1$, and that output is $X_1$, with the coordinates $P_1$, $X_1$ identifying a point on the firm's demand curve (point $A$ in the exhibit). The demand curve that this firm faces is identified by the alternative points that would be realized if the firm were to raise its price by varying amounts, or were to lower its price by varying amounts. Very critically for the model, the array of price and output values that could be realized by the firm as it changes its price depends on the strategic reactions of the other firms in the oligopoly.

Two alternative reactions, taken together, are the basis of the kinked demand curve model. When the firm we are focusing on changes its price:

1. The other firms match the price change exactly: a perfect response.
2. Or the other firms do not change their price at all; they keep their own prices constant: no response.

Which assumption do you think is the better of the two?

Two demand curves that pass through point $A$ are shown in Exhibit 6.7. The more elastic curve labeled $d^{NR}$ is the *no response demand curve*, and the less elastic curve labeled $d^{PR}$ is the *perfect response demand curve*. The different elasticities of the two curves reflect different market share behavior. With regard to the $d^{PR}$ curve, there is no reason to presume that a price change in either direction that is matched by all of the other firms in the industry will change market share, and so we assume that our firm's market share is constant along the $d^{PR}$ curve.

The $d^{NR}$ curve is more elastic than the $d^{PR}$ curve because if one firm raises its price and the other firms do not follow, that firm is now the high-priced competitor and, accordingly, it loses market share. Or, if the firm lowers its price and the other firms do not follow, that firm is now the low-priced competitor and, accordingly, it gains market share. Because the firm loses market share when it raises its price and gains market share when it lowers its price, its no-response demand curve is more elastic than its perfect response demand curve.

Think about it again. Which assumption, and therefore which demand curve, do you think best characterizes the situation that our representative oligopolistic firm is in? Here is the thinking behind the kinked demand curve. Each assumption, and thus each of the two curves, has a role to play. For a price increase, the no-response curve is relevant: if the firm decides to raise its price, its competitors will respond with the thought, "let it, let it be by itself with a high price and lose market share." But for price decreases, the perfect response curve is relevant: If the firm decides to lower its price, its competitors respond with the thought, “We had better match the decrease instead of being ourselves the high-price competitors and losing market share.”

Because of this asymmetric response on the part of the competitors, our firm’s demand curve is kinked, as we show in Exhibit 6.8. The kink is at point *A*, the point where we switch from one demand curve to the other. Above point *A*, the no-response curve is a solid line (to show that it is the operative curve), while below *A* the no-response curve is only dashed in (because it is not operative in this region). Above point *A* the perfect response curve is only dashed in (because it is not operative in this region), but below point *A* it is a solid line (because, for price decreases, the perfect response curve is operative).

The no-response and perfect response marginal revenue curves ($MR^{PR}$ and $MR^{NR}$) that are associated with the two demand curves are also shown in Exhibit 6.8. As with the demand curves, the relevant portion of each marginal revenue curve is a solid line and the nonoperative portion is dashed in. Because of the discontinuous drop in demand elasticity that occurs when we move down the demand curve past point *A* (the point of the kink), the operative part of the marginal revenue curve is discontinuous at the output $X_1$. The gap in the operative marginal revenue curve is the vertical section of the curve between points *B* and *C* in Exhibit 6.8.

The major implication of the kinked demand curve model follows from the discontinuous portion of the marginal revenue curve. As for any other firm, the best profit output for our oligopolistic competitor is at the point where marginal revenue equals marginal cost. Two marginal cost curves are shown in Exhibit 6.8: One is labeled $MC^{MAX}$, and the other is labeled $MC^{MIN}$. For the first, $MC^{MAX}$, the best profit output is $X_1$. For $MC^{MIN}$, the best profit output is also $X_1$. More generally, for any marginal cost curve that passes through the gap in the marginal revenue curve, the best profit output is $X_1$. The curve $MC^{MAX}$ is the highest marginal cost curve that passes through the gap, and $MC^{MIN}$ is the lowest marginal cost curve that passes through the gap.

The conclusion to be drawn from this model is that the oligopolist’s marginal cost curve can fluctuate considerably without affecting its best profit output and pricing decision. It is also possible, costs constant, for demand to shift back and forth (that is, to make parallel shifts, to the left or to the right) without impacting the equilibrium price (although these demand shifts will induce output changes).[3] The implication from this is that the administered price of oligopoly firms will not change with every fluctuation of their cost curves and/or demand curves. This lack of responsiveness is not only attributable to friction (the cost of changing an administered price), but also to the competitive actions and responses that can characterize oligopolistic competition.

---

[3] For there to be no price changes, the demand shifts must be parallel so that the height of the kink (which identifies the constant price) does not change.

## Duopoly[4]

You may remember the movie, *A Beautiful Mind*, starring Russell Crowe. Crowe played the role of John Nash, the brilliant economist/mathematician who won a Nobel Prize for having developed noncooperative game theory. His contribution to game theory—the Nash equilibrium—was a significant effort at demonstrating what happens when desirable outcomes are sought by players in noncooperative games. In a Nash equilibrium, no one player can gain an advantage by unilaterally changing his or her strategy if the other players do not appropriately coordinate their strategies.

Nash's basic insight was that when people compete, they do not necessarily achieve the most desirable results, results that would be obtained if they instead cooperated. The model is not based on any assumption of investor irrationality. Rather, every trader is assumed to be rational, to be acting out of his or her own best self-interest. Nevertheless, at the end of the day, each participant winds up with an undesirable outcome. That is the nature of a Nash equilibrium.

John Nash applied his classic idea, the Nash equilibrium, to a game known as the Prisoners' Dilemma. We explain it with reference to Exhibit 6.9. Assume that two prisoners, prisoner A and prisoner B, have both been accused of a minor crime and also of a serious crime, that they both expect to be convicted for the minor crime, and, very importantly, that they also expect to be acquitted for the major crime if neither of them admits to having committed it. The two prisoners are interrogated in separate rooms so that neither knows what the other has revealed or may reveal. They are found guilty (or not) for the major crime, and are punished (or not) depending on what each of them says when interrogated. If they both keep quiet, they each get a two-year sentence for the minor crime; if they both squeal they each get a five-year sentence; if one prisoner squeals and the other does

---

[4] Material in this section draws on research being undertaken by Chakraborty, Pagano, and Schwartz, as presented by Pagano at the May 4, 2004, Baruch College Conference, *Electronic vs. Floor-Based Trading*, and published in "Getting the Trades Made," in *Electronic vs. Floor-Based Trading,* ed. J. A. Byrne, A. Colaninno, and R. Schwartz (Springer Science + Business Media, 2006), 2–11, with the kind permission of Springer Science + Business Media.

| | B keeps quiet | B squeals |
|---|---|---|
| A keeps quiet | A: 2<br>B: 2 | A: 10<br>B: 1 |
| A squeals | A: 1<br>B: 10 | A: 5<br>B: 5 |

Outcome: Both squeal!

**EXHIBIT 6.9** The Prisoners' Dilemma (a Nash Equilibrium)

not, the squealer gets only one year and the nonsquealer gets a ten-year sentence. The alternatives, for each, are either (1) keep quiet or (2) squeal and incriminate the other.

Now for the dilemma. Both prisoners have an incentive to deviate from the "do not confess" option. Each will fare better by squealing and incriminating the other if the other one keeps quiet. The "squealer" (say prisoner A) gets just one year while the other prisoner (say prisoner B) receives a far more serious punishment—a ten-year sentence. So A has an incentive to incriminate B. For the same reason, B has an incentive to squeal on A so that he can possibly go free in just one year, and in any case avoid a ten-year sentence if he does not squeal but A does.

Here is the way to assess it: If A squeals, B is better off if B squeals (he gets a five-year penalty instead of a 10-year penalty), and if A does not squeal, B is also better off if he squeals (he gets a one-year instead of a two-year penalty). Therefore, whatever A does, B is better off if he squeals. The same is true for A: whatever B does, A is better off if A squeals. Exhibit 6.9 shows this. Thus both prisoners squeal, and they both get five-year penalties instead of the two-year penalties that they would have each received if they had both kept quiet.

The bottom line is that both have the incentive to incriminate the other, even though each of them would be better off if they both kept quiet, and they both know all of this. And so the story has a sad ending (for the prisoners, at least). They both squeal. Consequently, the two prisoners wind up with heavy sentences and are in a worse situation than if they had cooperated (instead of having competed) with each other. But the police did get the crooks. With regard to the two prisoners, Nash equilibrium has enabled justice to be done.

Nash equilibrium can be applied far more broadly (that, after all, is what any model is for). Two firms embroiled in duopolistic competition can employ competitive pricing strategies (e.g., offer discounts), marketing strategies (e.g., engage in heavy advertising expenditures), and/or product differentiation strategies (e.g., have fancy and costly product packaging). A Nash equilibrium could result with regard to each of these strategies. The two competing duopolists could very well keep their prices higher and spend less on advertising and packaging if they were to operate cooperatively rather than competitively.

Another application of Nash equilibrium applies to the world of equity trading. In the next section, we look at a situation where one large, institutional investor is looking to execute a big buy order, another large, institutional investor is looking to execute a big sell order, and the two traders act competitively, not cooperatively.

## HOW COMPETITION PLAYS OUT IN THE EQUITY MARKETS

In this section, we turn again to the equity market to take a closer look at competition in a specific industry. First, we consider how large traders do or do not submit their orders at market openings so that a limit order book may or may not build (i.e., may or may not amass liquidity). The difficulty of building a book is that large institutional traders typically hold their large orders back—in whole or in part—in an attempt to mitigate the adverse price impacts that they can have—impacts that drive up investors' trading costs. Second, we turn to the dealer market. We begin by picking up a few themes discussed in Chapter 5 and giving further background on the equity dealer market, then proceed to a monopoly dealer market, and lastly turn to a competitive dealer market.

### The Large Traders' Dilemma[5]

In this subsection we consider the competitive (noncooperative) behavior of two large equity traders, one a buyer and one a seller, who are each considering whether to (1) submit a big order to a morning call auction or (2) in an attempt to control their market impact costs, send a flurry of chopped-up orders over a period of time to the continuous market that follows the call (we identified both the call and continuous market structures in Chapter 4).

---

[5] Material in this section draws on research being undertaken by Chakraborty, Pagano, and Schwartz, as presented by Pagano at the May 4, 2004, Baruch College Conference, *Electronic vs. Floor-Based Trading*, and published in the conference proceedings book by the same title, ed. Byrne, Colaninno, and Schwartz (Springer Science + Business Media, 2006).

Using a Nash-type game theory structure in a duopoly-type setting, we obtain an insight into how this strategic decision is made, and an understanding of the dynamics of book building.

Let each trader know that the other one is out there (a best-case scenario), but each is uncertain about whether the other will show up at the opening call. Comparable to what we showed in the previous section for the prisoners' dilemma model, the optimal outcome for both large traders, if they could coordinate with each other, would be to go to the opening call. If they did, they would meet, provide liquidity to each other, and happily buy and sell with little or no market impact.

But as is the case with the prisoners' dilemma, both traders have an incentive to deviate from the mutually beneficial strategy. Let's call the buyer "A" and the seller "B." If A holds his buy order back and B enters a large sell order in the call, price will be pushed down, and A will be able to buy at a lower price in the continuous market that follows. The same logic holds for trader B: if A enters the call alone and pushes the price up, B will be able to sell at a higher price in the continuous market that follows. And so, for each of them, holding back (which is equivalent to squealing in the prisoners' dilemma story) while the other places an order in the call leads to an excellent result. On the other hand, not holding back when the other does, and going into the call alone, leads to a very bad result.

Consequently, as with the prisoners' dilemma, they both do something that is undesirable for the two of them collectively—they both hold back from the opening call. The result? Both traders go to the continuous market, which is less liquid, and they each wind up paying for their decision by incurring higher market impact costs. Nevertheless, A's behavior is not irrational, and neither is B's. Each is basically following a self-optimizing trading strategy, given each trader's expectations of what the other trader will do. Nevertheless, the Nash equilibrium outcome is not desirable for either of the large traders.

A market structure lesson can be learned from the application of Nash equilibrium to equity trading. A call auction is a highly desirable trading venue for participants if (and this is a big if) a sufficiently sizable number of traders submit their orders to it. But individually, large traders in particular can have a strategic incentive to avoid the call. Consequently, the book may not build adequately in the opening call, and the inherent value of this trading facility may not get realized.

The bottom line is that bookbuilding (which can be viewed as liquidity building) does not occur easily in the nonfrictionless environment. Market architects, when designing a call auction, should as much as possible strive to mitigate the costs associated with Nash equilibrium.

## Overview of Market Maker Operations

Equity dealer markets generally involve intense competition between a limited number of dealers: between roughly three dealer firms for a thinly traded, small-cap stock, and fifty (or more) dealer firms or more for a heavily traded, big-cap stock. For the most part, oligopoly is the competitive model that best fits the dealer market.

As we have discussed in Chapter 5, the key economic function of a market maker firm is to provide immediacy to public buyers and to public sellers. In so doing, the dealer buys for its own inventory when public participants wish to sell shares, and sells from its own inventory when public participants are looking to buy shares. In providing this service, the market maker, of necessity, accepts inventory swings; the swings involve risk, and carrying risk is a cost that must be compensated for. The classic compensation for dealer services is the bid-ask spread. Let's state it this way: the bid-ask spread is the price of dealer services.

Along with setting the spread, the market maker also determines the *level* of its bid and offer quotes. It does so to achieve reasonable price discovery for the stock it is making a market in and, very importantly, to control its inventory positions (as we have also discussed in Chapter 5). Picture a market maker who needs to reliquify after having acquired a long position in XYZ stock of 150,000 shares. He or she will adjust the quotes downward. By improving on the offer (posting a lower ask), the market maker is indicating a more aggressive willingness to sell shares. By posting a lower bid, the market maker is discouraging further public selling. Alternatively, a market maker firm with a large short position will raise its offer to discourage public purchases, and will raise its bid to encourage more public sales. But, after having "adjusted the tiller," the market maker, like the captain of a ship, must wait for the public to respond so that its inventory can be brought back to a reasonable level. Inventory control is like piloting a boat—steering is not always so simple, particularly at slow speeds, when response lags are more appreciable.

Market makers have a second inventory control mechanism—they can also trade with other market makers. This procedure is akin to steering a car. With a car on a dry road, the driver turns the wheel and the car responds immediately and exactly (in contrast to the boat mentioned above). It is common for one market maker who is long shares to sell to another market maker who has a short position and is looking to buy shares (and vice versa). The transaction is referred to as *interdealer trading*.

A dealer may also give price improvement. *Price improvement* refers to the practice of executing a market buy order at a price that is lower than the best posted offer, or executing a market sell order at a price that is higher than the best posted bid. Market makers get to know their customers. Consequently, they can differentiate between those who are apt to be trading for liquidity reasons and those who are likely to have come to the market because they are in possession of new information. Liquidity motivated customers include, for example, an indexer who is trading to rebalance its portfolio in order to track an index, or a mutual fund that is experiencing cash inflows or outflows from its investors. On the other hand, a value investor with the reputation of being a good stock picker may be trading because of better information. Dealers are more apt to give price improvement to customers who they believe are trading for their own individual, idiosyncratic reasons, rather than because they are in possession of private information that will soon become common knowledge.

Another common dealer market practice is called *preferencing*. Preferencing refers to a customer sending an order to a particular market maker regardless of what that dealer might be quoting at the time (or when the quote might have been posted). This can be done because there is no time priority rule in a quote driven market (with a time priority rule, orders posted at the same price execute according to the time when they were placed, with the earlier orders getting priority). Public customers (this applies primarily to institutional clients) are free to pick the market maker firms that they send their orders to, and their orders are typically preferenced to a market maker who in the past has developed a special relationship with the customer.

What if the market maker firm that the customer chooses to trade with is not posting the most aggressive quote? If its quote is less aggressive than the best quote in the market, the market maker is not obliged to take the order, but generally will accept it to maintain a good relationship with the customer. When a market maker does accept an order, he or she will typically fill it at a price equal to the most aggressive bid or offer that is existing at the time the order is received. For instance, if the best bid in the market is 50, a sell order preferenced to a dealer quoting a \$49.90 bid will be filled at 50. The practice is referred to as *quote matching*.

## Monopoly Dealer Model

With reference to Exhibit 6.10, we now analyze equilibrium in a monopoly dealer market. We let all orders be the same size (one round lot) so that the aggregate buy and sell propensities of public traders can be simply described by the number of orders that would arrive at the market per an interval of time. In Exhibit 6.10, the curves labeled $R_B$ and $R_S$ show the arrival rates for the buy and sell orders, respectively. The negatively inclined $R_B$ curve reflects the fact that, as the share price is lowered, more investors want to buy one

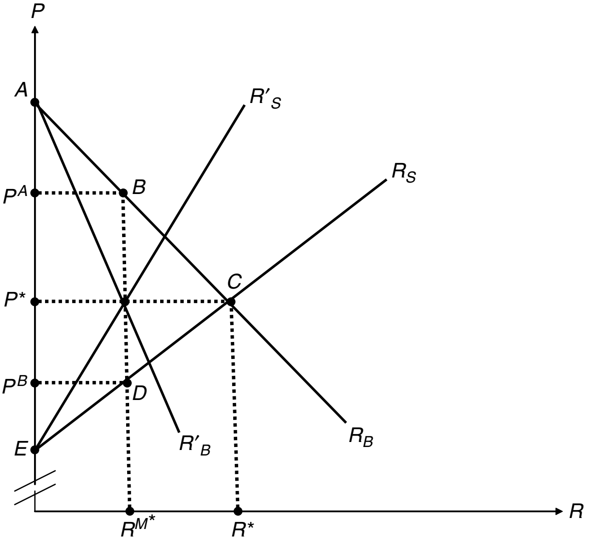


**EXHIBIT 6.10** Price and Volume Equilibrium for a Dealer Firm That Is a Monopolist with Respect to Its Sale of Shares to the Public, and a Monopsonist with Respect to Its Purchase of Shares from the Public

round lot, and the positively inclined $R_S$ curve reflects the fact that, as the share price is raised, more investors want to sell one round lot. Exhibit 6.10 is consistent with Exhibit 5.12 but, rather than simply establishing why a spread exists in a dealer market, it shows more explicitly how the dealer might set its bid and ask quotes for shares so as to maximize its profits.

In Exhibit 6.10, $P^*$ is an equilibrium price. At a price of $P^*$, the arrival rates for buy and sell orders both equal $R^*$. The arrival rates are the numbers of orders that are *expected* to arrive over some period of time (for example, a trading day); the *actual* order arrival is an uncertain (stochastic) process. Due to the sporadic nature of order arrival, however, the counterpart buys and sells that come to the market do not meet each other in time, the reason being that they do not arrive at precisely the same moments in time. Think about it: buy and sell orders can arrive at exactly the same time only when they each arrive continuously on the market (as they would if trading was a frictionless process). But if they arrive sporadically, they are bound to miss each other, if by only a micro second. Therefore, in the absence of any other market mechanism (such as a public limit order book), the dealer firm must exist for the market to function.

The monopoly dealer firm sets its quotes at levels that maximize the firm's expected trading profits. Because the industry comprises only a single dealer, the dealer is a monopolist vis-à-vis public participants who are looking to buy shares. As such, the dealer's demand curve is the downward sloping linear line labeled $R_B$ in Exhibit 6.10. The firm's associated marginal revenue function is the downward sloping linear line labeled $R'_B$ in the figure.

But to be able to sell shares to the public, the dealer firm must also buy shares from the public. Because only one dealer is making a market in the stock, the firm also has a monopoly position with respect to its purchase of shares from its customers. A single buyer, however, is not called a monopolist. Instead, the term used is *monopsonist* (and while an industry with one seller is a *monopoly*, an industry with just one buyer is a *monopsony*).

And so, the single dealer firm is a monopolist vis-à-vis its sale of shares to the public, and it is a monopsonist vis-à-vis its purchase of shares from the public. In Exhibit 6.10, the upward sloping curve labeled $R_S$ shows the rate at which the public will sell shares to the monopsonist. In this simplified setting, the upward sloping curve can be taken to be the monopsonistic dealer's average-cost curve. Associated with its average-cost curve is the dealer's marginal cost curve, which is the upward sloping curve labeled $R'_S$.

The dealer firm, like any other firm, maximizes (expected) profits by equating the marginal cost incurred when buying shares with the marginal revenue realized when selling shares. In Exhibit 6.10, the upward sloping marginal cost curve, $R'_S$, intersects the downward-sloping marginal revenue curve, $R'_B$, at a point that establishes $R^{M*}$ as the monopoly dealer's equilibrium expected rate of transactions. Accordingly, the dealer firm's optimal ask is $P^A$ (which is read from the $R_B$ curve), and its optimal bid is $P^B$ (which is read from the $R_S$ curve). These quotes and the associated quantity ($R^{M*}$) identify the expected profit-maximizing price of the monopoly market maker's service, and the amount of the service that, given its quotes, the firm will expect to provide.

The role of dealer inventories that we discussed in Chapter 5 can be revisited in the context of this model. The dealer has to accept acquiring both long and short inventory positions as trading progresses. In part, this is because of the *asynchronous pattern* with which buy and sell orders are transmitted to the market. Namely, buyers and sellers do not generally arrive in a neatly interspersed pattern (first a buyer, then a seller, next a buyer, once again a seller and so on). At times buyers will happen to arrive in bunches, and the dealer will be short shares; at other times, sellers will happen to arrive in bunches, and the dealer will be long shares. Even if purchases and sales tend, on average, to balance over relatively long periods of time, they commonly do not balance over relatively short periods. This means that dealer inventory positions will fluctuate, and these fluctuations bring risk into the picture.

There is a second factor that accounts for inventory fluctuations: The order flow is *stochastic*, and $R^{M*}$ is the *expected* rate (not the actual rate) of buy and sell arrivals. After the fact, it may turn out, for a trading session, that more orders in total have arrived to buy at the dealer's ask ($P^A$) than to sell at the dealer's bid ($P^B$). In such an event, the dealer firm, on net, has reduced its inventory of the stock, or has acquired a short position. Or perhaps more sell orders have arrived given the dealer's bid than buy orders have arrived given the dealer's ask; in this case, the dealer, on net, has reduced a short position or has added to its long inventory position. We considered these fluctuations in Chapter 5 with regard to the market maker's inventory policy.

We stress now, as we did in Chapter 5, that market making operations involve further costs for dealers. The quote setting solution presented in Exhibit 6.10 does not reflect the cost of risk that the dealer bears. The only two costs that are reflected in the exhibit are (1) the cost of buying shares so that they can in turn be sold, and (2) the cost of selling shares so that, in turn, purchases can be made. When the cost of risk is introduced into the analysis, a more complex model is required to solve for the optimal dealer quotes.

Further information can be extracted from Exhibit 6.10. With reference to the figure, we can identify the cost of friction in the marketplace to market participants collectively. By participants collectively, we are referring to public buyers, public sellers, and also the dealer firm. The friction that we are referring to is trading costs that keep public buy and sell orders from arriving continuously at the market, as they would in a frictionless environment.

Let's repeat something: Because they arrive intermittently, the public buy and sell orders do not meet each other in time. Therefore, a mechanism must be in place to handle the orders when they do arrive. As we have seen, the dealer, as an intermediary, is such a mechanism (the dealer firm sells with immediacy whenever a public buy order arrives, and buys with immediacy whenever a public sell order arrives). But dealer intermediation is not a free good—dealer operations are themselves costly. The costs incurred by the dealer are a manifestation of friction in the marketplace.

Let's return to Exhibit 6.10. In the figure, the intersection of $R_S$ and $R_B$, at point $C$, establishes $P^*$ and $R^*$ as the frictionless market equilibrium solution. If this solution were to be attained, the (consumer) surplus for public buyers would be the area $ACP^*$, the analogous (producer) surplus for public sellers would be the area $P^*CE$, and the gains from trading for the public participants collectively (both buyers and sellers) would be the area $ACE$.

In contrast, in the dealer-intermediated, nonfrictionless market, public customers buy at the dealer's ask ($P^A > P^*$) and sell at the dealer's bid ($P^B < P^*$). At these prices, the market clears with $R^{M*}$ transactions, on expectation, taking place. The dealer earns $P^A - P^B$ for each round trip, which translates into an expected total profit of $(P^A - P^B)R^{M*}$. With this information, we can contrast the gains from trading in the nonfrictionless market with the gains realized in the frictionless market by all participants collectively (public buyers, public sellers, and the dealer).

In the nonfrictionless dealer market, (consumer) surplus for the public buyers is the area $ABP^A < ACP^*$, (producer) surplus for the public sellers is the area $P^BDE < P^*CE$, and profits for the dealer is the area $P^ABDP^B$. The sum of these gains is given by $ABDE$, an area that is less than the collective gains to trading in the frictionless market, which is the area $ACE$. The difference between the total gains to trading in the frictionless market and in the nonfrictionless market is $ACE - ABDE$, which is the area $BCD$.

We have seen this result before with reference to Exhibit 6.2, where the area $ECF$ is the deadweight loss for the market. In Exhibit 6.10, the area $BCD$ is also a deadweight loss. The loss of trading surplus for public buyers and sellers is partially offset by the market maker's profits. But, over and beyond this redistribution of gains, there is also a deadweight loss to all of the participants in aggregate. That loss is the area $BCD$ and, as we have previously said, it is a monetary measure.

In Exhibits 6.2, 6.6, and 6.10, the deadweight loss is a reduction in the gains from trading that is attributable to a theoretically desired equilibrium solution being unattainable. In Exhibit 6.2, the loss is attributable to the imposition of a sales tax; in Exhibit 6.6 it is attributable to a seller having some market power over price; and in Exhibit 6.10, the loss is attributable to trading being a nonfrictionless process. We see in the context of this model how the cost of operating in a real-world micro market reduces the benefits from trading that participants in aggregate would share in a frictionless environment. We also see that this communal cost can, in theory at least, be quantified.

## Competition in a Competitive-Dealer Market

We have considered three market structures: the first two, the perfectly competitive market and the monopoly market, are polar opposites, and the third one, imperfect competition, may be thought of as an intermediary case. The competitive dealer market is an oligopoly that fits into this third category—it is characterized by a relatively small number of firms that compete intensely with each other, that react to each other's strategic moves, and that take each other's competitive actions and reactions into account when formulating their own strategies.

A curious thing about the competitive dealer market is that, as a rule, no one dealer firm, alone, sets the market bid-ask spread, which is the price of immediacy for public customers. This is because the market spread is rarely the spread of an individual dealer. Rather, the market spread, which is also called the *inside spread*, or the *inside market*, is the difference between the lowest ask of all of the dealer firms and the highest bid of all of the dealer firms. Generally, one dealer firm (or more) sets the lowest ask, and some other dealer firm (or more) sets the highest bid. As we just said, it is rare for any one dealer to be at both the best market bid and the best market ask at the same time.

Price competition takes place among dealer firms as the market makers compete with each other by how aggressively they set their bids and offers. If a market maker firm raises its quotes while keeping its spread constant, its quotes are more aggressive on the bid side; if it lowers its quotes while keeping its spread constant, its quotes are more aggressive on the offer side. The only way to become more aggressive on both the bid and the offer simultaneously is for the dealer firm to narrow its spread. But dealer firms rarely do this. That is why it is rare for one dealer firm at the same time to set both the best market bid and the best market ask for a stock.

Competing by aggressive quote setting is not a simple matter, however. The dealer market practice of preferencing is a big reason why. Preferencing diminishes a market maker's incentive to compete via the aggressiveness of its quotes. While a market maker firm that is quoting at the best bid or offer has a somewhat higher probability of receiving the next incoming order (especially if it is alone at the quote), the next order could nevertheless be preferenced to a market maker who is not on the inside market. And so, with preferencing a common practice, what does a market maker firm accomplish by raising the market's best bid, or by lowering the market's best offer? It will have raised the bid or lowered the offer that the other market makers will have to match, and it may not itself even receive the next order. So, the incentive to make this quote adjustment is not high unless the market maker has very good reason to believe that the quotes currently posted on the market are out of line with the broad, underlying desire of public participants to buy and to sell shares. However, we do emphasize that reasonably accurate price discovery is important for the dealer firms. First and foremost, for inventory control purposes, market makers certainly want their quotes to be positioned appropriately, and they will change them if the quotes appear to be unreasonable from a price discovery perspective. That said, a market maker's incentive to compete by aggressive quote setting is relatively weak.

If market makers have only a weak incentive to quote aggressively, how do they compete? They compete by developing good customer relationships so that they will fare well in the battle for preferenced order flow. An institutional investor will call a specific dealer house because it has received good service from that firm in the past. If a dealer either turns a customer down or executes the customer's order at an inferior price, the customer will think twice before preferencing an order to that dealer firm again.

Market makers offer an array of ancillary services that enable them to attract order flow. For instance, they may provide customers with research reports on companies, and/or computer software, and/or data for investment analysis. They may offer direct computer links that result in faster executions than customers could achieve elsewhere. In addition, they may take their customers to fancy restaurants, ball games, and championship tennis matches. And so forth.

Because of the way in which market makers compete, inside market spreads (the lowest offer minus the highest bid) tend to be wider in dealer markets than in order-driven, limit order book markets. In most electronic order-driven markets, incoming orders execute against contra-side orders that have been selected according to two strict criteria: price priority and time priority:

**Price priority:** The most aggressively priced order (the highest bid or the lowest ask) executes first.

**Time priority:** If two or more orders are tied at the most aggressive price, the order that has been placed first executes first.

In an order-driven market, if there is a lengthy queue of orders at the best bid or offer, a newly arriving buy (or sell) order can get priority simply by being priced one tick above the best bid (or one tick below the best offer). Aggressive pricing to get ahead of the queue narrows the spread in an order-driven market. Because this does not occur in a quote-driven environment, spreads tend to be wider in dealer markets.

What effect does preferencing have on market makers' profitability? As we have noted, more orders will be preferenced to a market maker firm that has good customer relationships. This is highly desirable for the firm. For one thing, a dealer firm's net revenue is positively related to its trading volume (all else equal, of course). Additionally, a firm that sees a larger percentage of the order flow has an advantage with regard to price discovery, and thus can set its quotes more knowledgeably.

Preferencing, however, is not an unmitigated good. It is excellent for a dealer firm to receive preferenced orders, but preferencing makes inventory control considerably more difficult. As we have seen, with preferenced order flow, posting the *most* aggressive quote on the market does not assure that a market maker will receive the next incoming order, and quoting *less* aggressively than the best bid or offer does not insure that the market maker will not receive the next incoming order. We have likened controlling inventory through quote changes to steering a boat, rather than a car. With preferencing, the boat is being steered with a broken rudder. The difficulty of controlling inventory can negatively impact a firm's profitability.

Under which regime do market makers prefer to operate, one with preferencing, or one with strict time and price priorities? Preferencing is an industry practice and dealers are comfortable with it or they would not have continued the practice. Nevertheless, preferencing is a two-edged sword.

### The Size of the Minimum Tick and the Price of Dealer Services

Operations for dealer firms are more profitable when, all else equal, these firms can charge higher prices for their service. For the dealer firms, the price for their provision of immediacy is the inside market spread. Accordingly, dealers like inside market spreads that are relatively wide. We have already noted that spreads tend to be wider in a quote-driven market (that does not operate on the basis of time priority) than in an order-driven market (where time priority is enforced). We have also noted that, with preferencing, the incentive is not high for a dealer to narrow the market spread by posting a higher bid or a lower offer. Despite these two factors, support for yet wider spreads is desirable from the point of view of the dealer community. There is a way to put a floor under the minimum size of the spread, and it can be very effective: Impose a substantial minimum price variation (that is, impose a large tick size).

Spreads in a competitive dealer market are expected to be tighter than they are in a monopoly dealer market because, as is typically the case, competition drives prices down. But the spread price of dealer services can only be driven down so far. The tightest that any spread (any individual dealer spread or the inside market spread itself) can be is the minimum price variation, the tick size. This floor under the price of dealer services was perceived by some influential government policy makers to enforce the imposition of an unjustifiably high price for public customers. And so the industry was pressured to reduce the tick size.

In 2001, under pressure from the U.S. Congress, stocks in the United States changed from trading in sixteenths of a dollar (6.25 cents) to trading in pennies. A few years before that, the pricing increment for stocks had been reduced to a sixteenth from its long-standing value of an eighth of a dollar (12.5 cents). The change to penny prices has had a big impact on equity market operations in general, and on equity dealer market operations in particular. We return to this government intervention in Chapter 8.

The move from sixteenths to pennies did two things: (1) The industry switched from fractional pricing to decimal pricing (while traders had been totally comfortable with fractional pricing, many of us find decimal pricing easier to understand), and (2) the tick size was reduced. The public policy goal in reducing the tick size was to lower the minimum size of the bid-ask spreads that interdealer competition could drive the inside market down to. Government's overarching objective in lowering the spreads was to reduce the cost of trading for public participants. We return to this issue in Chapter 8.

In the good old days (for the dealers) when the tick size was $^1/_8$th of a dollar, dealer quotes tended to cluster on price points that were at even eighths (two, four, and six eighths and an even dollar). In other words, a quote might be seen at 23, or at $23^1/_4$, or at $23^1/_2$, or at $23^3/_4$, but not often at $23^1/_8$, or at $23^3/_8$, or at $23^5/_8$, or at $23^7/_8$. This practice received some justification from arguments such as it is easier for people to think in terms of quarters than in eighths. Regardless, the practice of skipping the odd eighths ($^1/_8$, $^3/_8$, $^5/_8$, and $^7/_8$) effectively put an even higher floor under the size of dealer spreads. These wider spreads were profitable for dealers. This pricing convention accomplished for the dealer firms what a pricing agreement for a cartel such as OPEC can achieve for the oil-producing firms. Let's put it this way: In the United States, market making is a less attractive industry to be in now that stocks are being priced in pennies instead of in eighths of a dollar.

## IMPORTANCE OF PRICE AS A COMPETITIVE VARIABLE

Whether an industry is perfectly competitive, monopolistically competitive, an oligopoly, or a monopoly, price is a variable of central importance. Consumption and production decisions are made with regard to price. Firms that are efficient, low-cost producers, can sell their products at lower prices and gain powerful competitive advantages in the marketplace. The relationship between price and marginal cost indicates how successful a market is in producing socially desirable levels of output ($P = MC$ indicates perfect output efficiency). But, when it comes to competition, price is not the only variable that matters, and price competition is not the only game in town.

In industries where output is not homogeneous and therefore is not fungible, firms compete intensely in a variety of ways. Product differentiation, innovation, the nurturing of good customer relations, and advertising are high on the list. Well-structured product lines and product bundling can bestow strong competitive advantages. A firm's reputation for delivering quality merchandise, the support that it gives its customers, the backing that it gives its products and the speed with which it provides these services are of critical importance. So too is the speed with which it can bring new products to market. Also, in reality, much serious time, effort, and funds are spent lobbying government to pass favorable legislation.

As much as possible, firms want their products to be branded, not commoditized. A good brand distinguishes a product, sets it apart, and lowers the perception that close substitutes exist. In the tradition of oligopoly and monopolistic competition, a good brand bestows some monopoly power on the firm that owns it. Successful branding lowers the elasticity of a firm's demand curve, thereby enabling the firm to charge higher prices. Many firms have well-recognized brandings: Coca-Cola, McDonald's, Best Western Hotels, and John Deere tractors, to name a few. Some people refer to tissues as "Kleenex."

For decades, the New York Stock Exchange enjoyed a distinctive branding. The Big Board was viewed by many as the biggest, strongest, highest-quality exchange globally. Two things in particular gave the NYSE its distinctive brand: the Exchange's trading floor and its self-regulatory operations (we discuss self-regulatory operations in Chapter 8). But self-regulation has now been separated from the Exchange's trading operations, and the Exchange's trading floor, while still in existence, has been largely replaced by electronic trading. Still big, still powerful, the NYSE is facing intense domestic and global competition. The Big Board is fighting to retain its distinctive branding.

Along with keeping its fees low, the Exchange competes by the speed of executions that it delivers, the quality of price discovery that it can achieve, the flow of information messages and orders that it can handle, the overall reliability of its systems, its product line (which now includes derivatives as well as equities), and its global reach (having merged in April 2007 with one of Europe's biggest markets, Euronext, the Big Board is now called *NYSE Euronext*).

This is a taste of what competition involves in micro markets. Competition is intense, it is complex, it lies at the heart of business operations, and it is global. Competition is an integral part of the nonfrictionless environment. While operational frictions, transaction costs, blockages, and other impediments create problems that must be dealt with, they also open opportunities that can be exploited by clever, intelligent, and fast-acting management. The big question that we all face is: How efficiently does competition play out in the nonfrictionless world? How close does it bring us all, as a society, to where we ideally would want to go? We address the issue of market efficiency in the next chapter.

## CHAPTER SUMMARY

Robust competition is the force that economists rely on to discipline the micro markets, to drive the markets to greater efficiency, to deliver product innovation, and to keep markets growing. How does competition play out?

What are the key competitive variables? How does competitive behavior depend on the structure of an industry? These are among the questions that we have addressed in this chapter. Here are the highlights:

- We first considered the perfectly competitive environment. Under perfect competition, all firms produce an identical (fungible) product and the number of firms is big enough so that no one firm is large enough to have any discernable influence on aggregate market outcomes. In the perfectly competitive industry, all firms are atomistic.
- So, how do perfectly competitive firms compete? Not in the ways that we typically think that firms fight it out in the marketplace. Rather, these firms collectively put pressure on each other via the effect that their aggregate output has on price. *Ceteris paribus*, higher prices lead to profits, profits attract new entrants into an industry, and the larger aggregate output that follows puts downward pressure on price until the excess returns are eliminated. Conversely, lower prices lead to losses that force firms out of an industry. In general, competition drives each firm to produce efficiently (i.e., to operate at a least cost combination of inputs and at a best profit output level of production).
- In long-run equilibrium, firms in a perfectly competitive industry do not make economic profits. If all firms in an industry are identical (that is, have identical production functions and cost curves), they each realize just enough total revenue to cover their costs of doing business. If the firms are not identical, those that are intramarginal (the firms with lower average cost curves) do realize excess returns, but these returns are called economic rents, not profits.
- Frictionless, long-run equilibrium has excellent economic properties for a perfectly competitive industry. Under perfect competition, price and industry output are determined by the intersection of the industry's aggregate demand and supply curves. Price (which we have seen in Chapter 2 is related to marginal utility) is equated with marginal cost, and both producer and consumer surplus are maximized. When the free and frictionless market equilibrium is perturbed by, for instance, the imposition of a sales tax, a deadweight loss is incurred by the industry's producers and consumers in aggregate.
- If perfect competition is on one end of the spectrum, monopoly is on the other. With perfect competition an unlimited number of firms are in the industry; with monopoly, the industry has just one firm. For both the perfectly competitive industry and the monopolized industry, we have very clear models of how price and output are determined in a frictionless, deterministic (rather than stochastic), and static (rather than dynamic) environment.

- Before an industry can be classified as being monopolized, the product under question (and therefore the industry) has to be defined. Is the product tubular steel, or is it steel in general, or is it all structural building materials? The answer depends on the cross elasticities of demand. If the cross elasticities are low for a particular kind of steel, than that particular kind of steel can be taken to be "the product," and if there is only one firm producing it, then that firm would be a monopolist. If the cross elasticities are high, then close substitutes exist and the product should be defined more broadly. Perhaps more than one firm is producing the more broadly defined product, in which case the single firm that is producing the specialized steel had best not be thought of as a monopolist.
- Like the perfectly competitive firm, the monopolistic firm maximizes its profits by producing at an output level that equates marginal revenue and marginal cost. But unlike that of the perfectly competitive firm, the demand curve facing the monopoly is the demand curve facing the industry, and the industry demand curve is downward sloping (that is, it is not infinitely elastic). In this case, marginal revenue is less than price, which means that the monopoly firm does not equate price with marginal cost. Rather, price is higher and output is less. This comparative static is undesirable from a public policy perspective. We have noted that "keep the markets competitive" is a mantra for many economists and government regulators.
- Several things were concluded about the monopoly firm: (1) It does not have a supply curve, and the very terminology does not apply (although its price and output are nevertheless determinant); (2) the monopoly firm will always operate in the elastic part of its demand curve (for the competitive industry, equilibrium is wherever demand crosses supply, and nothing prevents the crossing point from being in the inelastic portion of the industry demand curve); and (3) a monopolistic firm, if it is able to segment its market, can increase profits by price discrimination (selling less and charging a higher price in a less elastic segment, and selling more and charging a lower price in a more elastic segment).
- If monopolies are undesirable (at least in a static environment), why do they exist? The answer is three words long: "barriers to entry." Product innovation, patents, copyright protection, well-kept secrets, and a government license or a franchise are important barriers.
- Another particularly effective barrier to entry is economies of scale. A market is said to be a *natural monopoly* when with scale economies a firm's average cost curve is falling over output ranges that are large relative to the market's aggregate demand for the firm's product. We also

considered two related concepts, economies of scope and network externalities.

- Having completed our discussion of the two ends of the spectrum of market structures, we turned to the vast majority of firms that operate somewhere in between them in a rather grey environment called *imperfect competition*. Perfectly competitive firms cannot affect the prices of the products that they produce, and thus they do not set price. With imperfect competition and monopoly, the firm does set the price of its product. Thus it is an *administered price*. As we have noted, we do a good job modeling the price administered by a monopolist. But the models of imperfect competition are not as definitive, and we have a number of different formulations. That is why we just called the environment "grey."
- The first case of imperfect competition that we dealt with is *monopolistic competition*. A substantial number of firms operate in a monopolistically competitive industry but, unlike under perfect competition, the products that they produce are to some extent differentiated. Accordingly, the demand curve facing each firm in a monopolistically competitive industry is downward sloping. But because there are many firms in the industry (there are no serious barriers to entry), competition eliminates profits for each of the firms, much as it does for perfectly competitive firms.
- Profits are eliminated because their receipt attracts new entrants into an industry. As new firms enter, the demand curves facing the existing firms shift down and to the left until, for the marginal firm, its demand curve is tangent to its average cost curve. Because the demand curve is downward sloping, the average cost curve is necessarily downward sloping at the point of tangency, and thus the marginal firm's average cost is not minimized as it would be under perfect competition. We viewed the higher average cost under monopolistic competition as a price that is paid by consumers because they are willing to pay for variety, because they accept paying for the product differences that distinguish monopolistic competition from perfect competition.
- The other case of imperfect competition is *oligopolistic competition*. Whether the firms that are in such an industry produce a differentiated product or a fungible product, does not matter. The distinguishing feature of oligopolistic competition is that the firms are few enough in number so that each reacts to the strategic decisions of the others. Very importantly, knowing this, each firm takes its competitors' reactions into account when making its own strategic decisions. Let's put it this way: One reality alone characterizes competition in an oligopolistic

industry—the firms take into account the interplay between their own competitive decisions and the responses of their competitors.

- We presented a standard oligopoly pricing formulation that delivers an interesting economic explanation of why administered prices are relatively constant over time (as they are generally observed to be in an oligopolistic industry): the *kinked demand curve model*. The model is based on two different competitive responses: (1) a firm's price increases are not followed, but (2) its price cuts are matched. This asymmetric response puts the kink in the demand curve, the kink translates into a discontinuity in the marginal revenue curve, and the discontinuity accounts for the relative stability of price (the *MC* curve can shift up and down through the discontinuous part of the *MR* curve, and the demand curve can shift to the right and to the left without changing the best profit output).
- We next turned to a duopoly (two competing firms), and to the equilibrium formulation of the Nobel laureate John Nash. Nash's game theoretic model is structured with regard to competitive strategies in general, and it can be applied not just to the firms' pricing policies, but also to their other competitive decisions (for instance, their marketing and product differentiation strategies). The key insight delivered by Nash equilibrium is that although the duopoly firms may each be following a self-optimizing strategy, both firms attain an equilibrium position that is not as desirable as the outcome they could have attained if they had acted cooperatively.
- With the oligopoly discussion in hand, we switched gears and considered how competition plays out in the equity markets. To this end, we first applied the Nash model to a setting where two large equity traders (a buyer and a seller) are deciding whether to enter their big orders in an opening call auction or in a continuous market that follows. True to Nash equilibrium, our two players both did something that is mutually undesirable: They both avoided the call auction and waited for the continuous market.
- Continuing with our focus on the dealer market, we gave an overview of market maker operations, presented a monopoly dealer-pricing model, and described some of the intricacies of interdealer competition in an oligopolistic market maker environment. Competition in this equity market clearly is complex and multifaceted.
- The chapter concluded with a discussion of the importance of price as a competitive variable. Quite clearly, although price is of critical importance, it is not the only competitive variable in the micro markets. We provided a taste of what competition involves in the equity markets. A more general question is: How much can we rely on competition to

provide the desired carrots and sticks that keep us, as a society, on the path of robust economic growth and performance? We turn to issues concerning the efficiency of the marketplace, and the efficiency of government regulatory intervention, in the remaining two chapters of the book.

## LET'S CONSIDER AN APPLICATION: WHY FIRMS EXTEND TRADE CREDIT

As noted in the chapter, competition is intense between firms that face downward sloping demand curves and thus have some power to set their own prices. The section on monopolistic competition contains a statement that applies to all imperfectly competitive and monopoly firms: ". . . (a) firm will strive to differentiate its product better, to advertise, and to have sales promotions, all in an attempt to establish a brand name that distinguishes it from its competition. Product quality and good customer relationships count heavily in this effort." There is something else that the above-mentioned firm will do: extend trade credit to its customers.

Nonfinancial firms commonly extend trade credit to their customers.* A typical set of terms is 2/10 net 30. These terms mean that a firm which is selling a good or a service is telling its buyer, "If you pay within 10 days you will get a 2 percent discount and, if not, the entire amount due has to be paid in 30 days." By extending trade credit, the seller is providing financing to the buyer. And so the buyer has obtained two resources, not one, from the seller: the good or service that is being purchased and short-term financial capital. We are accustomed to banks providing short-term financial capital to firms. But banks are not the only ones. With trade credit, nonfinancial firms are also providing short-term funding to each other. Why do they do this?

Two motives can be noted: a finance motive and a transactions motive. The finance motive exists because of the time value of money, because buyers who receive credit from a seller will be better customers. The transactions motive exists because trading is not a frictionless process, because it costs something to match the time pattern of payments with the time pattern of receipts, and trade credit enables the

*This application draws from Robert A. Schwartz, "An Economic Model of Trade Credit," *Journal of Financial and Quantitative Analysis* (September 1974), 643–657.

(*Continued*)

(*Continued*)

payments to be coordinated more efficiently. The finance and transactions motives both suggest that the demand curve faced by a firm for its product is higher and further to the right when the firm offers its customers trade credit.

With this by way of background, let's focus on a firm that is facing a downward sloping demand curve, and consider the relationships between the firm's credit policy, the location of its demand curve, and its product price. Ignore the discount period and concentrate only on the net period (that is, instead of the terms being "2/10, net 30," let them simply be "net 30"). Assume that the seller is a relatively large and established firm, and that its customers are a homogeneous group of smaller, younger firms.

The relevant variables are:

$P_L$ List price of the product

$r_S$ Relevant rate of interest for the seller

$r_B$ Relevant rate of interest for the buyers (assume $r_B > r_S$)

$P_S$ Present value of the list price received by the seller $[P_S = P_L(1 + R_S)^{-N}]$

$P_B$ Present value of the list price paid by the buyers $[P_B = P_L (1 + R_B)^{-N}]$

$N$ Length of the net period

$Q$ Quantity of the good sold to the buyer

The questions are:

1. With reference to the *transactions motive*, explain why the selling firm's list price demand curve shifts up and to the right when the firm extends trade credit to its buyers.
2. With reference to the *finance motive*, how does the rate of interest for the seller and for the buyers depend on the characteristics of these firms (e.g., their size)? For whom do you expect the rate of interest to be lower, the seller or the buyers? (Remember, we have assumed that the seller, relative to the buyers, is a larger, better established firm.)
3. With reference to the *finance motive*, explain why the selling firm's list price demand curve shifts up and to the right when the firm extends trade credit to its buyers.

4. Recognizing that the seller can gain knowledge of its buyers through its marketing and sales efforts, how do you think the seller's rate of interest, $r_S$, compares with a rate of interest that a bank would charge the buyers?
5. Assume $0 < r_S < r_B$ and $N > 0$. What does this imply about the relationship between the list price, the seller's present value price ($P_S$), and the buyers' present value price ($P_B$)?
6. Present three demand curves that the seller could use in determining its credit policy, pricing, and output decisions: (1) a demand curve for $N = 0$, (2) a list price demand curve for some $N > 0$, and (3) a present value to the seller of the list price demand curve for $N > 0$.
7. If extending credit to its buyers is profitable for the seller, what economic reality would ensure that $N$ has a finite optimal value?
8. Do you expect that the optimal value of $N$ and the optimal value of $Q$ are positively related or negatively related? Why?
9. Assume that economic activity is expanding, monetary policy is increasingly tightening, interest rates and credit rationing are increasing, and trade credit is being substituted for bank credit. What might this imply about the efficiency of financial markets and the efficacy of aggregate monetary policy?

## TraderEx SIMULATION: INTERMARKET COMPETITION IN THE EXCHANGE SPACE

Again we turn to the TraderEx simulation software that is available in *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course* to give you a further opportunity to work with some of the concepts we have developed in this text. In the current chapter we have considered how different firms compete with each other under a variety of market structures that range from perfect competition to monopoly. We have seen that, while of special importance, price is not the only competitive variable. In this simulation exercise, we would like to place you on the management team of an equity-trading firm that is competing with another equity-trading firm for public order flow. Your actions in the exercise will involve your trading shares. The questions we ask you to address involve the advice you would give your firm concerning how to profitably gain more order flow.

(*Continued*)

(*Continued*)

Here is the setting:

- You are on the management team of the National Stock Market (NSM). The Exchange has just one trading platform: an electronic limit order book. NSM is facing increasing competition from MegaPipe (MP), an alternative trading system that delivers block trades to its institutional customers.
- MP is a dark pool trading venue (the price and size of orders in the system are not displayed to the market) that enables institutions to trade with each other whenever their orders match or cross in price, as long as the price is at or between the quotes posted on the public limit order book of the National Stock Market, where the stock is listed.
- Being on the management team at NSM, you want to experience trading in both venues to get a better understanding of how each system operates, to identify the advantages that each system offers, and to think more clearly about your firm's competitive responses to MP's growing success. Fortunately, you can trade in a simulated environment and not put real capital at risk.
- Start each of your simulation runs with a single objective: purchase 2,000 units of a stock in the course of one trading day. Assess your purchase prices against VWAP, the volume-weighted average price that the stock has traded at over the course of the day.
- Employ three different strategies: (1) Your trades are made entirely on NSM's open limit order book, (2) your trades are made predominantly in MP's dark pool, block market, and (3) your trades are made in roughly equal proportions in both venues. For each of these three strategies, run the simulation enough times to get a feel for how your average buying prices compare with VWAP for each of them.

Please respond to the following:

1. Assess your VWAP performance for each of your trading strategies (mainly NSM trading, mainly MP trading, and a fairly even balance of the two). Which strategy worked best for you?
2. No commissions were charged for trading in either venue. If commissions had been charged, would you have directed your order flow differently if the commissions were not the same in the two venues?
3. Assume that each of the two trading venues is charging a 2¢ a share commission and that the order flow is price-sensitive to the commission rate. You are considering lowering your commissions in an attempt to

capture more order flow. What do the kinked demand curve model and the Nash equilibrium model suggest would be the consequences of this price change? Present your own thoughts about the desirability of lowering your commissions.

4. What did you like and what did you not like about trading in each of the two venues?
5. How do you feel as an NSM executive about having MP as a competitor? Clearly the alternative market can attract some order flow away from the Exchange but, on the other hand, MP's presence might have some positive effects. What do you think? (Remember, large orders are not easily traded in large blocks on an open limit order book platform.)
6. Because MP has been gaining appreciable order flow, NSM would certainly like to take steps that would make it difficult for MP to gain more market share. What steps might you consider taking?
7. Do you realize that MP is free riding off of NSM's prices? Can you explain why? Can you explain what this means about the maximum market share that MP may attract away from NSM?
8. How does competition plays out in the order driven equity markets? How much do you feel we can rely on competition to provide the desired carrots and sticks that keep the equity markets on the path to robust economic growth and performance?

## CHAPTER 7

# Market Efficiency

In the previous chapters we have seen how consumption and production decisions are made in micro markets. While the discussion touched on issues concerning the efficiency of market outcomes, efficiency was not our primary focus. Now it is.

In this chapter, we first consider the ability of a frictionless market to achieve allocation efficiency (goods and services being produced and consumed in the right amounts from a public policy point of view). We then set forth ten classic causes of market failure (that is, failure to achieve an efficient allocation of resources). Next we turn to the information sets that we use to formulate our expectations and that we base our decisions on.

The chapter considers five dimensions of informational efficiency:

1. Efficiency with regard to an existing information set.
2. Efficiency with regard to information-gathering and information-producing activities.
3. The informational accuracy of equilibrium prices.
4. The dynamic efficiency of information dissemination.
5. The informational accuracy of market prices.

We also examine an empirical test of market efficiency and present a more comprehensive view of the efficiency of a financial market. Next we turn to a consideration of how market failure contributed to the financial meltdown that started in August 2007 and extended through 2009. We conclude the chapter by raising questions concerning the greater generality of the frictions problem that, to a large extent, underlies market inefficiencies.

The better part of this chapter on market efficiency focuses on information. Information, after all, is a driving force for the micro markets. Important issues concerning it that we address include the classification of information, the nature of expectations that are formulated given existing information (we explain the difference between homogeneous expectations and divergent expectations), and the informational accuracy of market prices. In essence, we treat information as a factor of production.

Economists have traditionally focused on three factors of production: land, labor (including human capital), and physical capital. Throughout parts of this book we have, without classifying it as such, added a fourth factor: financial capital. After all, desired combinations of land, labor, and physical capital cannot be easily obtained in a relatively illiquid, nonfrictionless market, and to acquire them financial capital is required.

Before proceeding, let's back up a minute and ask a critical question: what is a factor of production? A factor of production is a process facilitator (as distinct from an ingredient that becomes part of the output) for which there is supply, demand, and a market-determined price. Labor and physical capital, for instance, are facilitators in the production of cakes, while flour is an ingredient of a cake. By this definition, financial capital is also a factor of production. For instance, bank loans are demanded by firms and are supplied by banks; their prices are expressed as rates of interest; and the loans themselves are not ingredients in the outputs of the borrowing firms.

Closely associated with financial capital is liquidity. Vis-à-vis financial capital, liquidity is a more encompassing concept (it can be applied to any micro market), but it is also inherently abstract and intangible. Nevertheless, liquidity is necessary for production processes, it can be supplied (for instance, limit order traders supply liquidity to market order traders), it is certainly demanded (for instance, portfolio managers favor more liquid markets, all else constant), and it can be obtained at a price (for instance, the bid-ask spread is a price that is paid by market-order traders, and it is the price that is received by dealers and limit-order traders who supply immediate liquidity to the market order traders).

In this chapter, we pay prominent attention to information as a fifth factor of production (along with land, labor, physical plus human capital, and financial capital plus liquidity). The bottom line is that, in an imperfect, nonfrictionless environment, land, labor, and capital (physical, human, and financial) cannot be properly utilized without information (which is a facilitator) and the knowledge that information bestows. Information is definitely an input in production (and also consumption) processes, and it is by no means a freely available resource. Often we must find it, produce it, and/or buy it.

That said, information sets themselves can be enormous, incomplete, imprecise, and inaccurate, and inefficiency in the micro markets can, to no inconsequential extent, be attributed to information-related inefficiencies. A primary objective of this chapter is to explain in greater detail what all of this entails.

## ALLOCATION EFFICIENCY

Allocation efficiency means producing and consuming the right goods, in the right amounts, and with the right combination of the factors of production, where "right" means in *socially optimal amounts*. We have seen how allocation efficiency can be achieved when frictionless, perfectly competitive markets are in long-run equilibrium. In a simple two-good ($X$ and $Y$) and two-factor-of-production (labor and capital) setting, the following conditions must be satisfied for the socially optimal amounts to be realized:

$$P_X/P_Y = MRS_{X\cdot Y} = MU_X/MU_Y \quad \text{(for all consumers)} \tag{7.1}$$

$$\begin{aligned} P_{labor}/P_{capital} &= MRTS_{labor\cdot capital} \\ &= MP_{labor}/MP_{capital} \quad \text{(for all producers)} \end{aligned} \tag{7.2}$$

$$P_X = MC_X \quad \text{(for all producers)} \tag{7.3}$$

$$P_Y = MC_Y \quad \text{(for all producers)} \tag{7.4}$$

where $MRS_{X\cdot Y}$ is the marginal rate of substitution of $X$ for $Y$, $MRTS_{labor\cdot capital}$ is the marginal rate of technical substitution of labor for capital, and $MP$ stands for marginal productivity. These four equations tell us that equality must be achieved *on the margin*. If any one of these marginal conditions is not satisfied, a redistribution of resources in consumption and/or production is possible that would add to the total production of goods, and/or the total achievement of utility.

The desirability of having the marginal conditions satisfied can be illustrated with reference to the consumption decision. To this end, we focus, with reference to Exhibit 7.1, on equation (7.1). The equation says that the marginal rate of substitution of $X$ for $Y$ must be the same for all consumers of $X$ and $Y$. Let's pick two consumers and analyze their allocation decisions as if they alone comprised the market (note that we do not make market prices explicit in the analyses). Exhibit 7.1(a) shows four indifference curves for consumer A, and Exhibit 7.1(b) shows four indifference curves for consumer B. Give consumer A an initial endowment of $X_{A0}$ units of $X$ and $Y_{A0}$ units of $Y$, which enables the indifference curve labeled $U_{A2}$ to be achieved, and give consumer B an initial endowment of $X_{B0}$ units of $X$ and $Y_{B0}$ units of $Y$, which enables the indifference curve labeled $U_{B2}$ to be achieved. The two consumers, together, have $X_{A0} + X_{B0}$ units of $X$, and $Y_{A0} + Y_{B0}$ units of $Y$ that they are free to reallocate between themselves. We can see how the reallocation works with reference to Exhibit 7.1(c).

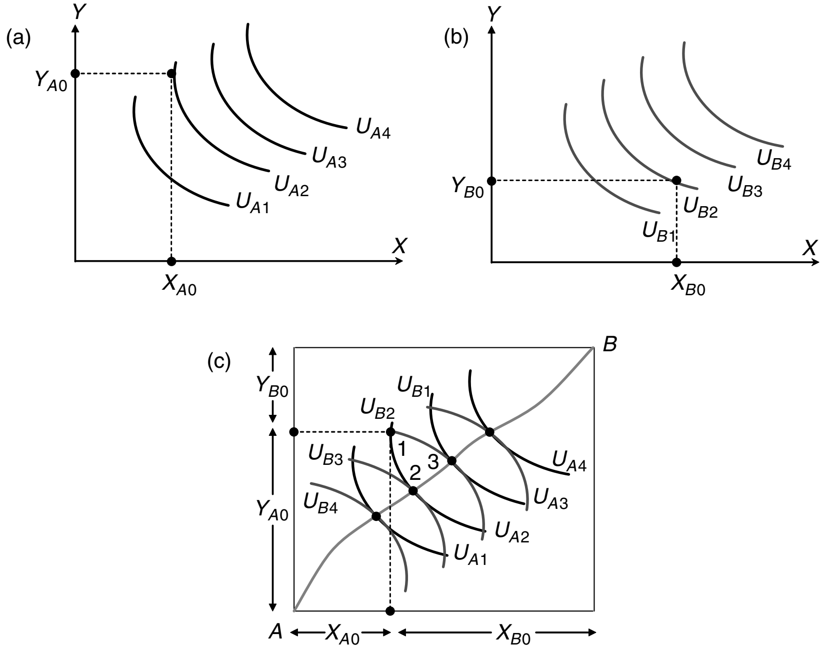


**EXHIBIT 7.1** Indifference Curves and an Initial Endowment of Goods *X* and *Y* for Individual 1 [7.1(a)] and Individual 2 [7.1(b)], and the Edgeworth Box and Contract Curve Formed by Uniting (a) and (b) [7.1(c)]

Exhibit 7.1(c) is obtained by combining 7.1(a) and 7.1(b). The two figures are combined by rotating 7.1(b) 180 degrees and uniting it with 7.1(a). When united, the two diagrams form a box called the *Edgeworth box*. The width of the Edgeworth box is the amount of *X* that the two consumers together have ($X_{A0} + X_{B0}$), and the height of the box is the amount of *Y* that they together have ($Y_{A0} + Y_{B0}$). The letter *A* at the lower left-hand corner of the box identifies this corner as the origin for consumer A; consumer A gains utility by moving to the northeast. The letter *B* at the upper right-hand corner of the box identifies this corner as the origin for consumer B; consumer B gains utility by moving to the southwest (this is because we have rotated the Diagram 180 degrees for B).

In Exhibit 7.1(c), the point labeled 1 shows the initial endowment of *X* and *Y* for the two consumers. Note that, at this point, the marginal rate of substitution of *X* for *Y* is not the same for consumers A and B because their respective indifference curves are not tangent to one another (and thus equation 7.1 cannot be satisfied). Accordingly, A and B will seek to trade with each other. By trading between themselves, A and B can attain any point in the box. Which point (or points) is optimal? Any move down and to the left would put A on a lower indifference curve; any move up and to the right would put B on a lower indifference curve. In fact, the only move that could put them both on a higher indifference curve is down and to the right through the area delineated by the $U_{A2}$ and $U_{B2}$ indifference curves. Because equation 7.1 is not satisfied, if properly directed, a move down and to the right would be utility enhancing for both consumers.

Here is one such move: slide down and to the right along A's $U_{A2}$ indifference curve from the point labeled 1 to the point labeled 2 in Exhibit 7.1(c), where A's $U_{A2}$ indifference curve is tangent to B's indifference curve which is labeled $U_{B3}$; the move enables B to reach a higher curve while leaving A on the same curve (and so A is not worse off while B is better off). Here is another move: slide down and to the right along B's $U_{B2}$ indifference curve from the point labeled *1* to the point labeled *3*, where B's $U_{B2}$ indifference curve is tangent to A's indifference curve which is labeled $U_{A3}$; the move enables A to reach a higher indifference curve while leaving B on the same curve (and so B is not worse off while A is better off).

We have drawn an upward sloping line that extends from A's origin (the lower left-hand corner), through points 2 and 3, and on up to B's origin (the upper right-hand corner). This line is the locus of all of the points of tangency between A's family of indifference curves and B's family of indifference curves. Thus, every point along this line satisfies the equation 7.1 condition that the $MRS_{X \cdot Y}$ be equal for all consumers (note that, for simplicity, we have left the price relative, $P_X/P_Y$ out of this analysis). Conversely, equation 7.1 cannot be satisfied at any point that is off this line. This line is known as the *contract curve*. Given the contract curve, here is a third move: Go to a point on the contract curve between points *2* and *3*; such a move would be utility enhancing for both participants.

Now look at Exhibit 7.1(c) once again and assure yourself that a move to the "proper region" of the contract curve (e.g., the region between points *2* and *3*) from any point that is either above the curve or below the curve would be utility enhancing for both participants. Also recognize that if the starting point (point 1) is located elsewhere, the "proper region" along the contract curve will shift.

We have just seen that a mutually beneficial trade will exist for consumers A and B if they are at any point that is not on the contract curve, and so it is to the contract curve that they will go. When starting from a point such as the point labeled "1" in Exhibit 7.1(c), they can go to any point on the contract curve from point *2* to point *3*. In other words, we have not obtained *a unique equilibrium*. Exactly where the two consumers will wind up between points *2* and *3* depends on their relative bargaining abilities (which is why the curve that passes through points *2* and *3* is called a "contract curve"). We could, of course, obtain a unique equilibrium by introducing a market-determined price relative ($Px/Py$) for the two goods, and by expanding the number of consumers (with just two isolated consumers, we can easily have a situation where their two marginal rates of substitution are equal but are not equal the price relative, $Px/Py$). Turn back to Chapter 2, Exhibit 2.5 and you will see an analogous formulation where equation 7.1 is fully satisfied.

There is one other thing. Achieving any point along the contract curve means that goods $X$ and $Y$ have been efficiently allocated between consumers A and B. The allocation is efficient given their initial endowments. As we have noted, if the initial endowment of $X$ and $Y$ between A and B is changed, the location of point 1 in Exhibit 7.1 is changed and so the locations of points 2 and 3 will change as well. The closer point 1 is to the lower left-hand corner of the Edgeworth box (A's origin), the wealthier B is relative to A, and the closer point 1 is to the upper right-hand corner of the box (B's origin), the wealthier A is relative to B. Recognizing this, the efficient allocation solution may not seem reasonable at all from a social welfare point of view. This, however, is an issue that we can point out but do not pursue further here.

## MARKET FAILURE

Assuming that markets are frictionless and that information is perfect justifiably simplifies much of our analysis, and it helps us to obtain important insights that would otherwise be unduly difficult (if not impossible) to get. But we must also take account of the fact that real-world markets are not frictionless environments, that various transaction costs and blockages (such as those we considered in some detail in Chapter 4) impede the making and implementation of optimal production and consumption decisions, and that information sets are rarely complete and perfectly accurate. In the complex environment of real-world markets, buyers and sellers have only imperfect knowledge regarding many goods and services and, as we will see later in this chapter, the complexity, magnitude, incompleteness, and imprecision of information further perturb the efficiency of market outcomes.

And then there is something called *market failure*. Failure may be partial (meaning that a market has failed to satisfy fully the optimality conditions that we will describe), or complete (meaning that a market has collapsed, that it has completely ceased to function). The ten classic conditions under which a market can fail are:

1. Imperfect and asymmetric information.
2. Irrational behavior.
3. Short-run myopia.
4. Externalities.
5. Network externalities.
6. Moral hazard.
7. Principal-agent problems.
8. Public goods.
9. Technological inertia.
10. Systemic risk.

## Imperfect and Asymmetric Information

We commonly act on the basis of information that is extensive in scope, complex, incomplete, and imprecise. Under these conditions, decisions may be made on the basis of expectations that are formed given the information at hand, or decisions may be postponed as additional information is gathered. Just how people behave in the face of imperfect information is a complex subject; we deal with it in further detail in the next section of this chapter.

Imperfect information commonly involves *asymmetric information*. That is, when information is imperfect, some participants have better knowledge than others. Whenever you deal with a professional service provider (for instance, a doctor, a lawyer, or a home builder), the professional will typically have considerably more knowledge than you do. How do you deal with this informational imbalance? By getting recommendations and second opinions, by developing an ongoing relationship with the service provider, and by doing some research on your own. Nevertheless, the asymmetry of information injects risk into a transaction.

A securities dealer, for instance, is always apprehensive about the possibility of buying from or selling to a better-informed public investor. The dealer typically does not know just why a customer wants to sell or to buy. Is the order motivated by news that the dealer has not yet received, or by a liquidity need, or by something picked up from technical analysis? Always, the market maker asks the question: is the customer in possession of some information that he, the dealer, does not know about?

A car dealer from whom you just bought a new car will ask the same question if you suddenly reappear wanting to sell the automobile fifteen minutes after you have driven it out of the showroom. What have you discovered about the car that I was not aware of, the car dealer wonders?

Could it be a lemon? Sure, it could be, and accordingly the market value of your new vehicle drops substantially the moment you take possession of it and drive away.

### Irrational Behavior

Market participants may behave irrationally. At times, emotions are involved when in the cold light of logic they should not be. Ignorance also plays a role. The sheer complexity of a situation can affect market outcomes. For instance, it is well established in the behavioral economics literature that changing the phrasing of a complex question without fundamentally altering the question's import can affect a respondent's answer and thus his or her behavior in the marketplace.

Irrationality can also express itself in terms of the intertemporal decisions that we make. That is, we might put excessive weight on the importance of current events and too little on the payouts (both plus and minus) that our current decisions might generate further in the future (this behavior is referred to as "short-run myopia").

### Short-Run Myopia

If a parent of a young child offers the son or daughter a cookie, their child will take it. What if the parent gives the child a choice: one cookie right now or two cookies in five minutes? Almost certainly, a fairly young child will take the single cookie right away. The kid will do this even though, at any conceivable rate of interest, the present value of two cookies in five minutes is clearly greater than the present value of one cookie right now.

Perhaps the child, with one cookie in hand, believes that with a bit of cajoling the parent will still fork the second one over in five minutes? Perhaps the child considers the two-cookies-in-five-minutes offer just too risky because the parent might get distracted by a phone call and forget the whole thing? Or, if young enough, perhaps the child does not even understand how short five minutes really is? Here is another explanation: the child is focusing so myopically on the current moment that the future might just as well not exist.

There might be a young child in all of us. Of course, as older, more mature people, we think of tomorrow and plan for the future. For sure, we discount future events for risk and uncertainty and, quite understandably, our risk-adjusted discount factors are larger for events that will occur further into the future. The question is, are our risk and uncertainty discount factors too high? Do we put too little weight on future events and too much weight on current results? In light of the young child in all of us, the answer might be yes.

Do we as a society pay too little attention to, for instance, threats to our environment (e.g., global warming) and to infrastructure development (e.g., to roads and bridges that need repair)? Do we as a society pay insufficient attention to investments in health and education that, at a cost to the current population, would benefit future generations? Some markets may equilibrate intertemporally but, in an imperfect, nonfrictionless environment characterized by incomplete knowledge, there is no assurance that all markets will do so acceptably well.

### Externalities

An externality exists when the exchange of a good or service between some participants in a market affects others who are not directly involved in the transaction. A classic example is the air or water pollution that can attend the production of goods such as aluminum and electricity. Pollution costs that are imposed on the broader economy are not taken into account in the output and pricing decisions (which explains the term *externality*) of individual consumers and firms, and thus inappropriate prices and socially inefficient output levels result.

Pollution and global warming are *negative externalities*: they impose costs on innocent third parties. Collectively, we bear the cost of bad air and polluted water, and collectively we attempt to control these adverse effects. But without full-cost pricing, some of us continue to drive gas-guzzling cars and persist in dumping waste into our waters because, for any of us individually, the negative externalities that such actions produce are costless.

A resource such as education provides *positive externalities* to the extent that better-educated people are better citizens (which one would generally expect). Yet, individually, we consider only how each of us might benefit from our own education as well as from that of our family members. For this reason, there is strong economic justification for collectively appropriating public funds to further the education of the broader population.

### Network Externalities

One type of externality is a *network externality*. We focused on network externalities in Chapter 6 but, in light of their importance, repeat some of the discussion here. The term "network" applies in a multiplicity of settings. For instance, the ability of people to communicate with each other by phone or by fax machine is an example of people being networked together by phones and by fax machines. Standardizing a software product such as Microsoft Word similarly allows people to send, receive, and use each other's files. Networks are generally more valuable to users the larger they are (the more people who own fax machines or who use Microsoft Word, the more people there are to whom you can send faxes and with whom you can share Word files).

A positive network externality is an economy of scale for the network (that is, a larger network is a better network). The term "externality" applies because individuals joining a network (e.g., by buying a fax machine) consider the value that they receive, but not the benefits that their participation brings to others because, for the joining individuals (who have created them), these benefits are an externality.

We noted in Chapter 2 that "as more orders converge in a marketplace, the orders provide more liquidity to each other and better price discovery results for the broad market." But customers as order-placers do not individually take into account the fact that, in a securities market, liquidity and price discovery are products of the network. The unfortunate truth is that Adam Smith's invisible hand fails to lead people to contribute optimally to the size of a network and liquidity, as a consequence, can be undersupplied in a securities market.

### Moral Hazard

A moral hazard situation exists when a decision maker, having undertaken an activity without bearing full responsibility for all possible outcomes, has an incentive to act more recklessly than he or she otherwise would and, as a consequence, imposes a cost on others. For instance, somebody who has rented equipment (perhaps a car) and has simultaneously purchased damage and theft insurance may not take adequate care of the equipment for the very reason that he or she is insured. Well-crafted contractual agreements can reduce moral hazard problems, but contracts typically do not cover all contingencies (i.e., are incomplete) and proper monitoring can be difficult.

Asymmetric information is part of the problem. A person who has rented some equipment has more information about the care with which it is being used than does the insurer who will bear the cost of any damage or theft. Consider, for instance, an out-of-towner who has rented a car and parked it on a city street, but who, not giving it proper attention, failed to lock the doors properly. Then the car is stolen. Who do you think knows better if the car was or was not properly locked, the out-of-towner or the car rental company?

### Principal-Agent Problems

Moral hazard translates into *principal-agent problems* when somebody (an agent) is acting on behalf of somebody else (a principal) and fails to act in the principal's best interest. A prominent example of a principal-agent relationship is the relationship between the owners (shareholders) of a firm and the firm's managers. A principal-agent problem arises because the managers' incentives may not be well aligned with the owners' objective of profit (wealth) maximization. For instance, profit maximization may be jeopardized by managers' having unnecessarily large offices, enjoying Friday afternoon golf games, and making dubious use of a company's private jet. How can the problem be dealt with? Principal-agent relationships are typically formalized by contractual agreements, but this in and of itself does not ensure that the contract terms will be followed. Monitoring is called for, but monitoring is costly, difficult to implement, and so it can deliver only limited protection for the owners' interests. Moreover, contracts are inherently incomplete (i.e., it is impossible to specify all contingencies). Another approach to resolving the principal-agent problem is to create an ownership interest on the part of the agents by, for instance, using incentive packages that include stock and stock options. Nevertheless, agents will continue to profit from nonmonetary payouts (i.e., private benefits such as larger offices and the use of a corporate jet) that are obtained at the expense of shareholders. And so principal-agent problems persist. They are notoriously difficult to resolve.

## Public Goods

For the most part, our discussion in this book addresses *private goods* as distinct from *public goods*. Bread, shoes, and haircuts are private goods. Somebody sells them, you buy them, and what you consume reduces the amount that is available for others. This is not the case with a public good. A lighthouse in the harbor is a classic example of a public good. A public park and military defense are other examples.

With all three (the lighthouse, the park, and military defense), there is just one quantity of the resource available and one person benefiting from it does not reduce the amount available for others (e.g., there is just one lighthouse in the harbor, and one ship picking up the signal does not prevent other ships from also being guided by the light). Moreover, nonpayers cannot be precluded from benefiting from the public good. That is why we pay for national defense through taxation, not via individual accounts as may be established with private security companies. The bottom line is that the private market does not produce optimal amounts of a public good. As its name implies, a public good must be provided by government.

## Technological Inertia

Inertia in a dynamic environment inhibits technological change. Inertia need not imply the recalcitrance of economic agents, but may be due to the cost of adopting a new technology. Consider, for instance, the introduction of an improved configuration of the keyboard. The standard keyboard is known as the *qwerty* keyboard because of the six keys on its upper left-hand side. An alternative keyboard known as *Dvorak* has been scientifically designed, is reputedly more efficient, but has not succeeded in the marketplace. The reason may lie in the following observations.

- Countless people currently know the touch system with the current ordering of keys. People acquainted with the qwerty keyboard would profit from learning the Dvorak keyboard only if the new system has a sufficient *differential advantage* over the old one, discounted over the remaining years for which they expect to be using the system.
- People just learning to type, on the other hand, will realize a return on their investment from the *total benefit* of knowing the new system. People just learning to type are also younger (on average) than those already acquainted with the existing system and hence they have more years to benefit from their knowledge (which increases their return on the investment).
- There are advantages to all typists knowing and using similar equipment. The benefit to any individual of knowing the new system may therefore be large if the market in aggregate converts to the new system, but small if the market in aggregate does not convert.

Consequently, the failure of existing typists to convert to the new system reduces the benefits of the superior technology to new typists, new typists therefore learn the older system, and the new and better keyboards are not produced, even though the social benefits in aggregate may outweigh the costs. This technological inertia is attributable to the inability of the new typists to compensate current typists adequately for their cost of learning the new technology.

Technological inertia is most likely to occur when change is comprehensive rather than partial. Comprehensive change affects an entire system (for instance, the full array of typewriter keys); partial change involves the components of a system (for instance, the introduction of new function keys on a keyboard). It may be possible in certain situations to achieve a comprehensive change by a series of partial changes, but such a switch in technology often is not possible.

Substantial technological change has taken place in the securities industry in the past few decades. Most prominent has been the introduction of electronic equipment to route orders and to disseminate information. Computer technology has also been applied widely to trade execution, but only after several decades of arduously slow adoption and in the face of much resistance from many participants. And trades for an array of financial instruments are still being made predominantly by person-to-person contact, either face-to-face or by telephone.

Computerized trading represents a comprehensive change from a human-intermediated system. The change required that professional traders adapt to an entirely new way of operating. Older participants whose style was suited to the noncomputerized way of trading feared that they would not fare very well in an electronic environment. Consequently, even after viable electronic systems were developed, technological inertia delayed their adoption.

New technology, of course, may eventually make its way in the market. The compact disc, for example, long ago replaced 33 rpm records, just as the 33 rpms had previously replaced the old 78s. But change takes time. A current challenge that we are facing is the green movement, which involves the introduction of solar and energy-efficient technology in cars, buildings, and elsewhere. Will green technology be introduced at a socially optimal rate? For this one, there appears to be insufficient assurance.

### Systemic Risk

An economy requires an infrastructure of, for instance, roads, rails, and financial markets. Parts of the infrastructure can be publicly owned (e.g., roads), but much is privately owned (e.g., railroads, securities exchanges, banks and brokerage houses, and even some roads). Widespread consequences can follow breakdowns in any of these markets. Regarding the financial markets, we have seen booms, bubbles, and busts, and, particularly after the stock market crash of 1929 and the 2008–2009 financial markets turmoil, extensive economic collapse. In the financial markets, breakdowns can occur due to technology failure or because of economic failure. Government involvement in regulating the financial markets has in good part been motivated by concern about the risk of systemic breakdown.

## INFORMATION AND EXPECTATIONS

Information is such a pervasive, integral driver of our decisions and therefore actions in the marketplace that informational efficiency is of paramount importance. What do we mean by the term? If you do not already know, you are about to find out. *Informational efficiency* is our primary focus throughout the remainder of this chapter.

When we consider the nonfrictionless environment, one property of the imperfect world is critical: the complexity, size, incompleteness, and imprecision of so much of the information that we base our decisions on (especially some of the most important decisions that we make). Of course, costs such as taxes, fees, transportation costs, travel costs, various slowdowns and delays, and so on also matter. Of course, blackout days, market closings, outages, a panoply of rules and regulations, and so on frustrate and stymie us. But put it all together and the most powerful of them all might well be the complexity, the vastness, the incompleteness, and the imprecision of so much of the information that we need to run our daily lives.

In the frictionless world, everybody has the information that they need. In that ideal environment, information is generated, distributed, and acquired at no cost and with no delay. Consequently, the second that it happens, everybody who needs it knows the news, understands it, and acts upon it. In the frictionless world, information is not imperfect, and imperfect information, therefore, is not a source of inefficiency.

Following George Stigler's pioneering 1961 *Journal of Political Economy* article, "The Economics of Information," information economics has played a major role in the economics literature.[1] As we turn our attention to information in this section, we structure the discussion with reference to the equity markets. The focus is specific and precise; the insights and implications that this focus yields can be applied with far greater generality.

Security prices depend on (reflect) investors' assessments of information. In fully efficient markets, information is reflected in prices with an accuracy that leaves no investor with any incentive to search for additional information or to trade. If information is perfectly reflected in prices and if trading is a frictionless (seamless and costless) process, then security prices will change randomly over time or, in technical parlance, a stock's price will follow a random walk (we explain what this means in a later section of this chapter). However, when the realities of actual markets are taken into account, it is clear that trading is not frictionless and that share prices do not follow random walks.

We will first focus on four dimensions of informational efficiency that are listed at the beginning of the chapter:

1. Efficiency with regard to an existing information set.
2. Efficiency with regard to information-gathering and information-producing activities.
3. The informational accuracy of equilibrium prices.
4. The dynamic efficiency of information dissemination.

---

[1] George Stigler, "The Economics of Information," *Journal of Political Economy* (University of Chicago Press, vol. 69, no. 3), 213–225.

Shortly we will deal with the fifth dimension, the informational accuracy of market prices (as distinct from equilibrium values). But before getting to any of these, we have several basic thoughts to put forth concerning information and the expectations that are based on information. In so doing, we pay primary attention to the securities industry.

## INFO ABOUT INFORMATION

Let us look at what lies beneath the information that is depicted by returns distributions. In actual markets, the "raw" information available pertains to market conditions and to the fundamental determinants of share value. To be useful, raw information has to be collected, processed, and analyzed. In light of the enormity of the available "information set," individual participants cannot be expected to know it all or to all process what they do have in identical ways.

Taking an aerial view, we classify security market information into two broad categories: (1) market (trading) information and (2) fundamental information that pertains to the investment decision (the basic determinants of share value).

*Market information* includes knowledge of the current quotes, last transaction prices, and transaction volume. In addition, some traders take account of recent high-low prices, the daily opening price, and the previous day's close. Furthermore, it would be of value to have information on orders that have not yet been executed, including knowledge of the limit order book, knowledge of orders held by floor traders at an exchange, and statements of buying or selling interest by block, institutional, and other large traders.

*Fundamental information* relating to the investment decision pertains to the determinants of future share values. For investors, the most useful form for fundamental information to take would be a statement of the means, variances, and covariances of security returns but, of course, these statistics are not readily available. A statistical approach to expectations formation would be to estimate expected means, variances, and covariances of returns from historic data (i.e., use retrospective empirical analyses of past stock returns). However, while history is important, expectations should be based on more than just past price behavior. And, be aware: returns distributions can always shift. Quite clearly, *current* information must also be taken into account. For this purpose, we now turn to fundamental information concerning a firm's future prospects. Here is what fundamental information comprises:

**Current financial information:** Information concerning current capital structure, earnings forecasts, and so on.

**Current strategy of management:** Knowledge of the strategic business decisions of management, which is based on management's assessment of current conditions and its outlook for the future.

**Current economic information:** Information concerning the firm's product market, the firm's competitors, national and international economic conditions, and so on.

**Structural change:** Knowledge of recent acquisitions, divestitures, discoveries, regulatory changes, and so on.

**Organizational efficiency:** Knowledge of corporate structure, managerial ability, and so on.

The five categories of fundamental information along with market information that we discussed previously pertain to the environment and to the firm whose security is being evaluated. One, however, might view information even more broadly. The relevant set encompasses attributes of the decision maker—the technical knowledge and experience that allow a good assessment of relevant facts. This information ranges from the decision maker's experience and skill at assessing intangibles such as managerial ability, to formal knowledge of portfolio theory and the capital asset pricing model. Information of this type may be nothing more than enlightened intuition; nevertheless, it is a key input into decision making.

Security analysis that is based on fundamental information involves the assessment of share value, but security analysts do not undertake a treasure hunt to find a golden number that one might call an "intrinsic value." Rather, as we have seen previously in this book, share prices are set the way they are for most resources—in the marketplace, in relation to the forces of demand and supply.

Information can be classified in another way—it can be characterized as public information, as inside information, or as private information:

**Public information:** Widely disseminated information that is readily available to the public. Being readily available does not necessarily mean, however, that the information is provided freely. For instance, real-time data on market quotes and transaction prices, along with many historic databases, have to be paid for.

**Inside information:** Information possessed by only a select set of people who have a special position with regard to the information. Corporate officers and others with business ties to a company (such as an investment banker, lawyer, or proofreader) are commonly in possession of "inside information." Inside information can itself be classified as either material or nonmaterial, depending on whether or not it would significantly affect share values if it was publicly known, or if before making an investment decision an investor would want to be in possession of the information.[2]

**Private information:** Information that investors may individually possess because of their own investigations and analyses of public information.[3]

Comprehensively viewed, the set of publicly available information is, in a word, enormous. Consequently, individuals see only parts of it and each participant typically assesses information in a way that, to some extent, is unique given his or her own vantage point. For this reason, private information plays a major role in the formulation of individual expectations. Tapping effectively into even a relatively specific information set can require considerable skill and expense, but an analyst who can do this successfully may earn handsome returns. An article in the *Wall Street Journal* on October 6, 2003 is illustrative.[4]

The *Journal* reported that, between June and August of 2003, certain large institutional investors using two detailed pharmaceutical databases which cost subscribers between $25,000 and $50,000 a year, were able to determine that the drug company, Schering-Plough, was losing market share for its hepatitis C medicine to a competitor, Roche Holding AG.[5] The investing public got hints of the development in July, but full information was not available until August 22. During this period, Schering peaked at $20.47 on June 18, and then drifted down, dipping below $17 as July turned into August. On August 21, the stock closed at $16.48. Strikingly, on

---

[2] For further discussion, see CFA Institute Code of Ethics and Standards of Professional Conduct, 9th ed. (2005).

[3] Employing a combination of public and nonmaterial inside information has been referred to as "Mosaic Theory" and, as stated in the CFA Institute Code of Ethics and Standards of Professional Conduct (op. cit.), there is no insider trading violation when "a perceptive analyst reaches a conclusion about a corporate action or event through an analysis of public information and items of nonmaterial, nonpublic information."

[4] Peter Landers, "Drug-Data Chasm Punishes Small Investors," *Wall Street Journal* (October 6, 2003), C1; and "Heard on the Street, Small Investors Face Handicap in Drug Sector," *Wall Street Journal Europe* (October 6, 2003).

[5] The data were "culled from thousands of pharmacies nationwide to give subscribers day-by-day tallies of the number and sales value of prescriptions filled."

August 22, the day when Schering confirmed the declining sales trend for its product, shares opened down $2.18 at $14.30. The *Journal*'s comment? "While nobody did anything wrong in this case – the investors weren't acting on inside information and the company violated no disclosure rules—it highlights an information chasm in the drug business that essentially punishes smaller investors." The large investors were acting on *private information*, and they were richly rewarded for their efforts. Undoubtedly, this story can be told many times over for many different industries.

### Homogeneous Expectations

Individuals form expectations about a company's future returns, given the current information set. Expectations link current information to the market value of shares. In our analyses, we can either assume that different investors have the same expectations concerning security returns (that is, have *homogeneous expectations*) or that they differ in their expectations (that is, have *divergent expectations*). In this subsection, we consider the former.

Even though the assumption of homogeneous expectations is known to be unrealistic, models based upon it (such as the standard Capital Asset Pricing Model [CAPM]) give much insight into how the market determines prices for various assets according to their risk and return characteristics. In the context of the CAPM, the assumption of homogeneous expectations is consistent with the assumption that decision makers have perfect information concerning the means, variances, and covariances of returns for the full set of stocks.

Rational decision making may seem to imply the homogeneity of expectations. This is because rational decision making considers what a rational person would conclude, given "the facts." Presumably, what one rational person would conclude, all rational people should conclude. However, having considered the elements that comprise the information set, we may better understand why the assumption of homogeneous expectations is unrealistic. It is plausible for a group of investors to have homogeneous expectations only if they share the same information set, process it in an identical way, and do not produce private information.

### Divergent Expectations

As we have said, the term *divergent expectations* is used to characterize an environment where investors' beliefs are not homogeneous. Individuals are commonly observed to differ in their assessments of information, and this is not difficult to explain. As we emphasized at the start of this section, information sets, being vast, complex, incomplete, and imprecise, are a challenge to understand. Different individuals possess only parts of the information that is publicly available, and some investors have private information. Furthermore, some individuals may also reassess their individual valuations based on what they come to know about what others are thinking. We refer to this as either *adaptive valuations* or as *interdependent valuations*. To illustrate, assume that the Fed announces in a depressed market environment that interest rates will be kept constant rather than lowered. A bearish expectation based on the announcement may be that the market will fall further because of the lack of a further monetary stimulus. A bullish expectation may be that the market will rise because the Fed does not believe that a further monetary stimulus is necessary.

Following the Fed announcement, put yourself in the shoes of a bearish participant, and consider your reaction when a slew of new buy orders arrive on the market and prices start to rise. Would you reconsider your own assessment of the Fed's decision in light of how others react? If so, you have adaptive valuations.

The divergence of expectations among investors along with the adaptive nature of valuations has major implications for the operations of a securities market. It explains why information change can lead to heavy trading. Namely, everyone does not react to new information identically and, as prices adjust to news, some choose to buy while others decide to sell. Further, in an environment where expectations are divergent, share values cannot be determined at the research desks of security analysts (and, as we will shortly see, shares do not have unique fundamental values). Rather, as is the case for most goods and services, prices must be set in the marketplace where the buy and sell orders of a large number of participants are brought together and translated into trades. In other words, *price discovery* occurs in the marketplace. As we will see in a later section of this chapter, price discovery is a complex, dynamic process. Recognition of this has major implications for the importance of having good market architecture.

## FOUR DIMENSIONS OF INFORMATIONAL EFFICIENCY

Having considered what information sets entail along with the expectations that are based on them, we now hone in on the first four dimensions of informational efficiency. Once again, here is the list of five:

1. Efficiency with regard to an existing information set.
2. Efficiency with regard to information gathering and information producing activities.
3. The informational accuracy of equilibrium prices.

4. The dynamic efficiency of information dissemination.
5. The informational accuracy of market prices (we deal with this one in the section, Price Discovery Efficiency, later).

How efficient are markets with regard to the first four dimensions? In a CAPM environment, the markets are perfectly efficient. In other words, in a CAPM world all participants are in equal and total possession of the information set (the market is fully efficient with regard to the first dimension), no information gathering activities are necessary (the market is fully efficient with regard to the second dimension), theoretically ideal equilibrium prices are obtained immediately (the market is fully efficient with regard to the third dimension) and, when new information arrives, it is disseminated to everybody instantaneously (the market is fully efficient with regard to the fourth dimension). But we do not live in a CAPM world and, accordingly, our discussion departs from that theoretically ideal, frictionless environment.

### Efficiency with Regard to an Existing Information Set

By an existing information set, we are referring to information that has been made publicly available and can be obtained by any participant and translated into a trading decision. Note that, in a nonfrictionless environment, not every investor actually possesses all of the information that is publicly available (information sets can be enormous, public postings are not necessarily seen by everybody, and some information has to be paid for). Nevertheless, if a sufficient number of investors have the public information and base their trades on it, the market can be fully informationally efficient with regard to its exploitation of the existing information set.

Investors make trading decisions in relation to existing information in two ways:

1. They search the publicly available information for situations where they think the market has mispriced an asset given the asset's risk, return characteristics. When such a situation is found, an investor takes a position in the asset that enables him or her to profit if and when prices are appropriately adjusted.
2. Even if all assets are appropriately priced, a menu of alternative mean-variance efficient portfolios exists, and the investor uses information concerning the risk, return characteristics of different securities to select an optimal portfolio (given his or her unique tastes for risk and return). Recognize that this would not apply in a CAPM world.

In this subsection, we focus on the first use to which information may be put.

The decision maker formulates returns expectations by assessing publicly available information, his or her own private information, and current market prices. If, on the basis of the assessment, the risk-adjusted expected return for an asset appears to be abnormally high (which means that the current price is abnormally low), the decision maker looks to buy additional shares. Alternatively, if the expected return seems abnormally low (which means that the current price is abnormally high), the decision maker looks to sell shares (if he or she is long in the asset), to short the stock, or simply to ignore the stock (if short selling is restricted). As you reflect on this, keep in mind that a negative return on a stock is a positive return for an investor with a short position and that, by shorting a stock, a trader who is bearish in relation to the market may also anticipate receiving positive returns.

Abnormal returns on an investment are, by definition, returns that are either higher or lower than an investor would require. For a bullish investor, abnormally high positive expected returns are exploited ("bought") by buying the shares, while for a bearish investor abnormally low positive expected returns or negative expected returns are exploited ("bought") by selling shares (if some are owned) or by shorting the stock (which, you will recall, means selling borrowed shares). Buying pressure increases current prices, which decreases expected returns to a long position. Selling pressure decreases current prices, which increases expected returns to a long position (and reduces expected returns to a short position). Because of the effect of purchases and sales on current prices, transactions that exploit the abnormal returns also eliminate them. It follows that abnormal expected returns are eliminated when investors achieve portfolios that are optimal, given the information set.

When all prices are properly set, the market is informationally efficient in terms of the first dimension. Namely, abnormally high returns (either positive or negative) cannot be realized by exploiting the existing information set.

### Efficiency with Regard to Information-Gathering and Information-Producing Activities

In the previous subsection, we considered the exploitation of a given information set. Information, however, is something that can be produced, and the size of the existing information set can be increased. (1) New public information can be *obtained* (gathered) by searching for it, and (2) new private information can be *created* (produced) by further analyzing existing public information. It is important to recognize that both activities are costly. We now consider how the information set can be expanded by information production and gathering activities. The second dimension of informational efficiency pertains to the optimal pursuit of these two activities.

Assume that an investor anticipates a particular mean and variance of returns, given the level of price at which the stock is currently trading. Also assume that, on the basis of logic and past experience, the investor has some understanding of how a new bit of information might alter the stock's price. At a cost, the investor can attempt to obtain that information before its impact is fully reflected in market prices. If successful, the investor benefits from any subsequent price adjustment that the new information might trigger.

With divergent expectations, an individual may also profit from information that has already been widely distributed to others. For instance, some information bit that is widely known can still be worth gathering for an individual if that individual were to interpret the information differently than the market. If that individual's assessment of the information turns out to be more astute than that of the market, then at some point in the future he or she will realize a positive return (even though the information bit has already been assessed by others and has had its impact on market prices). Thus information-gathering on the part of that individual would be justified.

As time goes by, truth reveals itself. As it does, some investors find that their anticipations were correct, and others find that they were wrong. Therefore, the return to information-gathering (and to information-production) includes the profits that one can achieve by being more correct than the market. This is indeed a difficult game to play, and few people believe that they can consistently play it with success. Nevertheless, security analysis is potentially valuable, even to decision makers who cannot consistently beat the market by being among the first to gather and produce the information.

Be sure to recognize how the second dimension of efficiency (efficiency with regard to information gathering and producing activities) differs from the first dimension (efficiency with regard to an existing information set). The first dimension addresses efficiency given a fixed information set: can I make profitable investment decisions based on mispricing given the existence of widely distributed public information? Implicit in this setting is the assumption that investors have homogeneous expectations. The second dimension addresses the question: can I realize enhanced returns by acquiring more information (either by gathering or producing it)? Implicit in this setting is the assumption that investors can have divergent expectations.

Obtaining/producing information along the lines just discussed is not the only way; a substitute exists. An investor can also *infer* informational change from market prices. That is, on the basis of past experience, an investor can interpret price changes as signals (albeit noisy signals) that some new information bit has been released. Therefore, rather than directly looking for the information, the investor may decide to let the price change signal the information.

We define four categories of investors with regard to their assessments of their own abilities to be more correct than the market:

1. **The rugged individualists:** Such people conduct their own security analysis, develop their own assessments of how a stock might perform in the future, and are totally unaffected by what others might think. For these people, future expected price levels are independent of a stock's current price.
2. **The sheep:** These investors exercise no independent judgment. Rather, they simply assume that the price set on the market is the correct price. This being the case, any price change is interpreted as signaling new information on the future value that share price will attain, but does not change expected returns. For these people the change in future expected price levels is proportionate to change in a stock's current price.
3. **The exaggerators:** These people assume that current price changes understate the impact of informational change. Thus they believe that any percentage increase or decrease in current price is associated with a greater percentage increase or decrease in future expected price levels.
4. **The rest of us:** Decision makers in this category think for themselves and come to their own conclusions, but also respect the market's collective judgment as it is reflected in security prices. These people revise their expectations of future price levels as current prices change, but do so less than proportionately.

Of these four categories, who produces and gathers information, and who does not? The rugged individualists and the rest of us (but less so) are producers/gatherers, while the sheep and the exaggerators are not.

What value might price signaling have to a member of the sheep category? From time to time new funds are invested in the market and old funds are withdrawn. Knowledge (or the belief) that the risk, return characteristics of a security have not changed is relevant for portfolio decisions made in relation to the injection or withdrawal of funds. In addition, the realization (or belief) that the risk, return characteristics of a security regain their previous values after the stock has adjusted to news may prevent the investor from mistakenly buying or selling after an opportunity has passed. In this regard, it may be advisable to act as a sheep if one does not have a preferential position vis-à-vis the information flow or special insight into information's meaning. Furthermore, the investor may in fact benefit from signals inferred from price changes. For instance, if prices do not adjust instantly and accurately to new equilibrium values, the investor may profit by quickly entering orders. Statistical tests have shown that some trading strategies based on past price changes do in fact generate excess gross returns (although transaction costs may make them, on net, unprofitable).

Technical analysts (chartists) in particular believe that profitable trading rules can be formulated on the basis of patterns exhibited by past price movements. Although charting is not accepted by many, the belief that the ebb and flow of investor reactions, psychology, and so on, introduce predictable, repetitive patterns is not, per se, erroneous. The reason for questioning the premise of the chartists is that in an informationally efficient environment, the exploitation of such price patterns would eliminate them. We consider this point further in a later section of this chapter.

In conclusion, information is valuable whether gathered/produced or inferred from prices, and whether obtained before or after the market has had a chance to adjust to it. However, when it is obtained sooner and directly (rather than later and inferred), it is (1) more valuable and (2) more costly to obtain. And so a trade-off exists in information-gathering/-production. The second condition for informational efficiency is that this trade-off is optimally resolved such that additional information-gathering/-producing activities do not generate abnormal profits. We next explain what this means. In doing so, we initially focus on optimality for an individual participant and then consider equilibrium for the market.

**Individual Optimality** Consider one individual's decision of whether to gather/produce (i.e., obtain) a single bit of information concerning a corporation. As discussed previously, the alternative for an investor who does not get the information is to make a portfolio decision regarding the corporation's stock on the basis of the stock's market price and other readily available information.

The value of the information bit to any specific decision maker depends upon the quality of the anticipations that he or she can formulate given that information, relative to the quality of the market's anticipations (which are reflected in the current price of the asset). The faster (relative to other market participants) the specific individual can obtain the information bit and the better he or she is as an information processor (relative to other market participants), the greater is the value of the information. The investor should look for information not yet obtained by others, and for new information that is not highly correlated with existing information.

Suppose a specific investor is moderately efficient at obtaining information. The larger the number of other investors who are informed and the more efficient others are as information gatherers and processors, the less likely it is that that investor would realize a competitive advantage by obtaining the information bit. If the decision maker's abilities are low enough in relation to the market, then he or she might do better by letting others obtain and process the information, and simply turn to price as an unbiased signal of the information.

The second condition for informational efficiency—that additional information gathering and producing activities do not generate abnormal profits—is equivalent to the requirement that an investor obtain information directly if its incremental value is greater than its incremental cost, or infer it from market prices if its incremental value is less than its incremental cost (including the opportunity cost of time).

**Market Equilibrium**[6] We have clarified that the value of additional information to each individual depends not just on the information itself, but also on that individual's efficiency vis-à-vis others at information gathering and processing. The market is in equilibrium with respect to a piece of information if all individuals for whom that information's value exceeds the cost of obtaining it, and if all individuals for whom that information's value is less than its cost infer it from the market price of the asset. In such an equilibrium, the information-obtaining activities that are undertaken are on net profitable, but additional information obtaining does not yield positive net returns.

To establish the existence of an equilibrium amount of information obtaining for the market, first consider a situation where no one actively seeks additional information directly, but where all participants base their expectations entirely on current market prices and on whatever information is publicly available. In this setting, the informational content of security prices would decline and prices would soon convey very noisy signals. In such a market, an investor with even a small amount of additional effort is able to spot some extraordinary situations—a stock paying a $15 dividend, offering much promise of dividend growth, and trading at $20 a share, or some other stock trading at $75 a share although the company has slashed its dividend and is about to collapse. In such an environment, additional information-obtaining activities are clearly profitable.

---

[6] For an expanded and more rigorous treatment of the material in this subsection, see S. Grossman and J. Stiglitz, "Information and Competitive Price Systems," *American Economic Review* (May 1976), used with permission.

What if many investors are informed? The informational content of security prices would then be high, and prices would convey a far less noisy signal. In this case, all but the most efficient information processors might find that the quality of prices set in the market is too good to beat. Accordingly, most people might simply follow the crowd rather than attempt to outguess the market. In the limit, if all share prices were to reflect all information fully and were not noisy signals, there would be no return to additional information gathering/production.

This may seem to imply a paradox: On the one hand, if stock prices were to reflect all information fully, no one would undertake security analysis; on the other hand, if no one were to undertake security analysis, stocks could not be appropriately priced.

There is no paradox. An equilibrium amount of information-gathering/-production exists. At one extreme, if virtually no one looks for information, the net returns to information-obtaining are likely to be positive for at least the most efficient information gatherers/processors. At the other extreme, if nearly everyone looks for information, the net returns to information-obtaining are likely to be negative for at least the most inefficient information gatherers/processors. When the market is in balance, an equilibrium number of investors are informed:

- The informed are those who are the most efficient at the process of obtaining information and/or those for whom information has the greatest value.
- For the marginal information obtainer, the value of the information just equals the cost of obtaining it.
- Those for whom the value of information is less than the cost of obtaining it infer the information from prices.

A market that has achieved such an equilibrium is informationally efficient with regard to the intensity with which information-obtaining activities are pursued.

### The Informational Accuracy of Equilibrium Prices[7]

As we have discussed in this book, prices act as constraints that lead individuals to use resources optimally, given supply and demand conditions. In traditional, deterministic economics, prices, being nonstochastic, convey no information about the resources because market participants are assumed to have complete information to begin with. On the other hand, when outcomes are uncertain and information is incomplete, prices play an important informational role. Prices are a mechanism for information transfer (from the informed to the uninformed) and for the aggregation of diverse information bits (for both informed and uninformed traders). In this subsection, we consider the efficiency with which prices perform these two functions.

---

[7] For an expanded and more rigorous treatment of the material in this subsection, see S. Figlewski, "Market 'Efficiency' in a Market with Heterogeneous Information," *Journal of Political Economy* (August 1978), used with permission.

With divergent expectations, a security's price reflects a weighted average opinion of all investors in the market. The more weight the market gives to the opinions of those who are better informed, the greater is the informational accuracy of the equilibrium prices. The third dimension of informational efficiency pertains to the quality of the equilibrium prices that are achieved.

Whose expectations might equilibrium prices reflect? People who believe themselves to be the most efficient at information-gathering and -processing are most likely to become informed traders (as we have previously noted, others will simply let price be their signal). With regard to the informed, two factors affect the dollar strength of each person's opinion, and hence the weight of his or her conviction in the market (the *dollar strength* of an investor's opinion is the funds that he or she commits to a position in light of the strength of his or her conviction). The first factor is the accuracy of the decision maker's opinion, and the second factor is the decision maker's wealth. Here is how they work:

**Opinion:** The dollar strength of an anticipation is correlated with the accuracy of that anticipation to the extent that truth carries its own conviction. Unfortunately, the presence of some bull- (or bear-) headed fools in the market makes the association between truth and conviction somewhat less than perfect.

**Wealth:** The wealth an individual has realized in the financial markets is his or her reward for having invested successfully in the past and, to an extent, the quality of a decision maker's earlier anticipations is correlated with his or her current abilities as a forecaster. Therefore, current wealth should be positively related to the accuracy of current opinions. However, this association is also less than perfect. Few are able to predict consistently well over time. Furthermore, some inefficient information processors may be richly rewarded simply by chance, and some efficient information processors may not do well—also by chance.

With regard to the informational efficiency of equilibrium prices, we conclude that even in a market that is informationally efficient in other respects, prices are noisy signals. We may never be sure who the most efficient information processors are by spotting the winners. Consequently expectations remain divergent and the market does not completely achieve informationally accurate prices.

### The Dynamic Efficiency of Information Dissemination

The three efficiency dimensions thus far considered concern static efficiency. We now turn to the fourth dimension, the dynamic efficiency with which information is disseminated in the investment community.

The efficiency of information dissemination has two dimensions:

1. The time needed for new information to be fully reflected in market prices.
2. The sequential order in which new information is disseminated among investors.

After change has occurred in a company's fortunes, investors should learn of it quickly, and the change should be quickly reflected in market prices. This is true for considerations of both equity and efficiency: market prices provide better signals to decision makers if they reflect current, rather than outdated, information. But because information dissemination is not instantaneous, some investors are bound to receive news before others.

On June 18, 1815, Napoleon was defeated at Waterloo. A carrier pigeon took news of the British victory to Nathan Rothschild in London. In a single day, Rothschild reaped a fortune by buying shares from uninformed, and quite frightened traders (he was also credited with having saved the London Stock Exchange). Rothschild's profit was not due to the news. It was due to his having received the news first.

A tremendous amount of information is disseminated in today's markets, and only nanoseconds may separate the receipt of news by different investors. Nonetheless, certain participants still receive information before others, and some may do so consistently. Thus we should continue to question the dynamic informational efficiency of the markets.

Investors receive information at different times for two reasons: (1) They are not equally efficient and aggressive as information gatherers, and (2) some people have a preferential position vis-à-vis the information flow. The efficiency and aggressiveness of investors should be rewarded, and they are. The returns to these efficient investors are the profits that they receive from the price adjustments which occur as others who have lagged behind them in the receipt and analysis of news come forward and trade. In part, it is the scramble to benefit from the price adjustments that accounts for the informational static-efficiency of a market. Nevertheless, the process of information dissemination should be fair.

On August 15, 2000, the U.S. Securities and Exchange Commission took a major step to promote full and fair disclosure of new information by adopting a new rule, Regulation FD. Reg FD addresses the problem of selective disclosure of information by requiring that, if a listed company provides any information to selected market participants who may seek to trade on the basis of that information, the company must also promptly disclose that information to the entire market.

Regarding preferential positions, for some people information is a by-product of a service they provide that has no relation to information gathering. The proofreader in a securities firm or a lawyer in a merger case, for instance, receives information that has not yet been released to the public. When these people profit from their prior receipt of information, we may observe certain proofreaders and lawyers being grossly overpaid for their services. No economic function is served by this overpayment. On the contrary, feelings of inequity on the part of those without preferential access can have harmful repercussions.

Insiders are deemed to have a preferential and unfair advantage vis-à-vis the information flow. Accordingly, these people are restricted in their freedom to trade shares of their corporation's stock. Insiders must file with the SEC after trading, and they are not allowed to trade on news that has not yet been made public.

A trade-off exists between the dynamic efficiency and the static efficiency of a market. The greater the flow of information in a market, the more accurate are the prices that are set and the more static-efficient is the market. But the flow of information is positively related to the return to information gathering, and the return to information gathering is in large part the price adjustments an informed trader benefits from when he or she receives the news first. Therefore, *dynamic inefficiency* (with some people profiting unduly at the expense of others) motivates the information-gathering activities that make a market more static efficient.

That such a trade-off exists is not surprising. Information gathering, like trading, is a manifestation of disequilibrium behavior. Also, like trading, information gathering helps to repair imbalances in the market and to bring prices back to equilibrium values. It is too much to expect that the dynamic process by which equilibrium is regained will generate no undesired side effects. If we want prices that are the best possible signals of information, we must let those who have the best information trade with a minimum of restrictions. Alternatively, if we do not want insiders to consistently exploit the uninformed public, we must settle for prices that are noisier reflections of the information set. This is indeed a difficult trade-off.

### The Informational Accuracy of Market Prices

It should be clear from the previous discussion in this chapter that a good deal is required for a market to be informationally efficient. But as if all of the above is not enough, there is a further consideration that concerns the efficiency with which prices are established in a marketplace: *price discovery efficiency*. Before getting to this, however, a few more concepts have to be identified, and we do this in the next section.

## A TEST OF MARKET EFFICIENCY

Let's return for a moment to our generic product, $X$. It is one thing to theorize about how the interaction of market forces produces market outcomes and how to assess the efficiency properties of these outcomes; it is another thing to test empirically the efficiency of the market for $X$. Sophisticated econometric techniques have been applied to the estimation of demand curves, supply curves, and cost curves, but this is never a simple, easy task. Of course, cost accounting itself is not simple, but light can be shed on the relationship between a firm's marginal cost of producing $X$ and the unit price it receives for selling $X$. The closer $P_X$ is to $MC_X$, the more efficient is the market for $X$.

One test of market efficiency is particularly compelling. It is an assessment of how completely information that gives stocks their values is reflected in share prices. The intriguing thing is that, without explicitly judging the relationship between stock prices and fundamental values, strong inferences about the efficiency of market outcomes can be drawn simply by assessing how stock prices evolve over time. This is an intellectually clever and compelling test of informational efficiency, and to it we now turn.

Let's start at the beginning. In the present context, the term *informational efficiency* refers to the accuracy with which prices that are set in a marketplace fully reflect the underlying information that these prices are based on. In an informationally efficient securities market, all existing information is fully reflected in share values, and no trader has an informational advantage that would pave the way to the realization of excess returns. The null hypothesis, which is referred to as the *Efficient Market Hypothesis* (EMH), is formalized as follows:

Excess returns cannot be realized from information that is contained in:

- **Past prices:** This is referred to as *weak form efficiency*.
- **Public information (including past prices):** This is referred to as *semi-strong form efficiency*.
- **All information:** This is referred to as *strong form efficiency*.

We focus on the first item, the test of weak form efficiency. The test addresses one concept in particular: whether or not stock prices follow *random walks*.

Random walk is a cool concept. In a nutshell, it means that where you take your next step is totally independent of all of the steps that you have previously taken. That being the case, anyone following your path would have no ability whatsoever to predict where your next step might be taking you. You would be myopically focusing on just the present step. Your friend who is trying to guess your destination will find that the path that you have thus far traced out contains absolutely no useful information. It is as if your footprints did not exist.

The idea of a random walk has been applied to the stock market. Those of us with an interest in the equity markets like to follow prices. We like to predict where they might be going next. Is a stock heading up? Is it drifting down? Has it gone up too far (or down too low)? Is it about to change direction and revert back to an earlier value? Or, perhaps, is price, like the directionless person in the previous paragraph, only following a random walk? If our guess is correct, unless you have taken (or are currently taking) a finance course that deals with the subject, you likely think that stock prices following a random walk is a preposterous idea.

It is not. In fact, many financial economists are of the opinion that stock price movements are essentially free of any correlation pattern. For the most part, they believe that, aside from transaction prices bouncing between bid and offer quotes, it is not possible to predict where a stock's price might be heading next. Our colleagues would be correct if markets were fully efficient, frictionless, and perfectly liquid. Why is that?

The answer starts with something that we have been saying in this chapter: information is an input that drives investment decisions and trading. Security prices are based upon the information set. As we have stressed, informational efficiency means that security prices fully incorporate the entire information set. If this were not the case, some clever participant could grab hold of the information set and churn out some profitable trades. If someone could do this at some moment in time, then the price of a stock, at that moment, could not be reflecting the information set correctly. If this were the case, the EMH would not hold.

So let's presume that a stock's price, at any moment in time, does properly reflect the full information set. What then would cause the stock's price to change? Any future price change cannot be attributed to the current information set. If it is, then the stock's price could not have been an appropriate reflection of the current information in the first place. For the EMH to hold, a stock's price can change only because of the advent of *new* information (news). The news would have to be totally unanticipated because any *anticipation* is part of the current information set. Now for the punch line—if the news must be totally unanticipated, its effect on the stock's price cannot be predicted. Therefore, neither can any future price change be predicted. The stock's price must, therefore, follow a random walk.

But what if it doesn't? If share prices do not follow random walks, then their path into the future offers some predictability, which means that profitable decision rules based on recent price movements could possibly be formulated. Using recent price behavior to predict future price movements is called *technical analysis* (another term, which is not used as much as in the past, is *charting*). A more recent tool called "algorithmic trading" is put to much the same purpose. While any decision rule can be called an algorithm, *algorithmic trading* refers specifically to computer driven trading rules. When an *algo* that is resident in a computer program is activated by, for instance, an intraday price pattern, a trading decision is formulated and an order is shot into the market, all within tiny fractions of a second (milliseconds, or less).

Both technical analysis and algorithmic trading are widely used by stock market professionals although neither would have a payoff if share prices followed random walks. And so, despite their being widely used, many economists discount their value. In part the economists' pessimism on this score is founded in the widespread belief in the efficiency of markets. One prominent economist, Burton Malkiel, wrote in 1973 that

> *Technical analysis is anathema to the academic world. We love to pick on it. Our bullying tactics are prompted by two considerations:* (*1*) *the method is patently false, and* (*2*) *it's easy to pick on.*
>
> —Burton Malkiel, *A Random Walk Down Wall Street* (New York: W.W. Norton & Company, 1973), p. 116

What do the data say about this? Going way back, before the 1970s, random walk tests have been among the most important tests of the EMH. For years, the EMH generally passed the tests and the EMH appeared to hold, at least in its weak form (namely, that all information in past stock prices and quotes is fully reflected in current prices). The results were certainly sobering for any young buck poised to predict future price movements based on patterns that he or she seemingly saw in historical data. To this day, the EMH should keep us all acceptably humble. But this does not mean that prices strictly follow random walks, or that correlation patterns do not exist, or that technical analysis and algorithmic trading do not have valid roles to play.

With the advent of powerful computers and high frequency data (complete, moment to moment records of all transaction prices, quotes, and transactions volume), financial economists are increasingly finding that stock returns are not entirely free of the correlation patterns that random walk theory says should not be there. And yet the EMH remains compelling. What is the missing element? Missing is full recognition of an important reality—markets are not fully efficient, they are not frictionless, and they are not perfectly liquid.

## PRICE DISCOVERY EFFICIENCY

We can now bring the fifth efficiency consideration into the picture: the efficiency with which prices are actually discovered in the marketplace.[8] Various different approaches to price discovery are possible, depending on the competitive structure of an industry (perfectly competitive, imperfectly competitive, or monopoly) and on the architectural structure of the marketplace (an auction market, a bargaining/negotiated market, a market where prices are administered, and so on). Perhaps the product in question is a Picasso painting being sold at auction where an auctioneer calls out ever higher prices until only one bidder remains. Perhaps the product is treasury bills sold at an auction where sealed bids are submitted in advance of the auction. Perhaps the product is cans of chicken noodle soup on the shelf of a neighborhood supermarket (if the cans are not selling, the administered price is lowered or, if they are flying off the shelf, the price is raised).

Here is another one, and it is our focus in this section of the chapter: price determination for shares of a stock. We ask the following question: after an overnight nontrading period during which important company-specific news and/or macro news has been released, how is price actually found the following morning when trading resumes? We answer this question in the context of a simple, plain-vanilla market structure: a continuous-trading, limit-order book market. But before getting to the answer, we have a story to tell.

---

[8] Our analysis in this section draws from Jacob Paroush, Robert A. Schwartz, and Avner Wolf, "The Dynamic Process of Price Discovery in an Equity Market," *Managerial Finance*, 2010, forthcoming.

Our story is about investor behavior that underlies the dynamic process of price formation. It recognizes that investors are not perfect information-processing machines, and it suggests how a set of investors might interact with each other as they collectively establish share values in the marketplace. In so doing, the story provides a conceptual foundation for technical analysis and algorithmic trading.

## Investor Behavior

The starting point of our story is that investors are not textbook efficient information processing machines. No one is able to assess share values with great precision. As we have been saying in this chapter, the problem is that information sets are too big, they are too complex, they are too incomplete, they are too imprecise, and our tools for analyzing them are too crude. Who can say, in light of a company's latest earnings report, dividend policy, investment plans, regulatory situation, competitive environment, management team, and the broader economic health of the country, and so on, that a company's equity is worth precisely $45.01 a share? Nobody can do this, not even a top graduate from one of the world's best business schools.

Let's restate the question. Given all the relevant information, could any stock analyst say that a company's expected growth rate is exactly 7.000 percent? What if one analyst sets the expected growth rate at precisely 7.000 percent and another one, who is equally precise but somewhat more bullish, sets the expected growth rate at 7.545 percent? It is ridiculous to believe that anyone could estimate a growth rate with basis point accuracy. Not even close! But let's carry on with the story.

Take these two growth-rate estimates and plug them into a simple stock valuation model, the commonly taught dividend discount model, and see what you get. Suppose that, one year from now, the company will pay a dividend of $1.35 a share, and that its required rate of return (ROR) on equity is 10 percent. With the dividend discount model, the current share value (Price) is

$$\text{Price} = \frac{\text{Dividend One Year from Now}}{\text{ROR on Equity} - \text{Expected Growth Rate}}$$

Plug in the numbers. If the expected growth rate is 7.000 percent, we have

$$\text{Price} = \frac{\$1.35}{0.10 - 0.07000} = \$45$$

If the expected growth rate is 7.545 percent, we have

$$\text{Price} = \frac{\$1.35}{0.10 - 0.07545} = \$55$$

Wow! As the growth rate tiptoed up from 0.07000 to 0.07545, the price jumped from $45 to $55!

The dividend discount model is, of course, a gross simplification of reality. Nevertheless, we can learn something from it. A future dividend of $1.35, a required rate of return on equity of 10 percent, and an expected growth rate of 7.000 percent translate into a share price of $45.00. Increase the growth rate by just 54.5 basis points to 7.545 percent, and the share value leaps 22 percent to $55.00. What this model says is that anyone who misestimates a growth rate by as little as 0.5 percent can evaluate the shares of this stock with only ±22 percent precision! This "fuzziness" in share valuation is the start of our story.

The second part of our story deals with how investors respond to fuzzy share valuations. Ultimately, each investor comes to his or her own share value. Each, in other words, sets his or her own reservation price. In the context of the example presented above, let's suppose that some investors (the gloomy bears with an expected growth rate of 7.000 percent) select $45.00 as their valuation, and that others (the brave bulls with the 7.545 percent expected growth rate estimate) select $55.00 as their valuation. What we have is a case of *divergent valuations* (which we discussed earlier in this chapter). In this story, the divergence of opinions is attributable to the inability to assess share values with precision. At little loss of generality, we keep the story simple by assuming that all investors fall into just these two groups: the gloomy bears and the brave bulls.

The third and final part of our story is that investors, recognizing that their abilities to assess information are limited, communicate with each other and respond to what others think. People talk at cocktail parties and in meetings. They listen to the often-conflicting opinions of the market gurus, and they observe the behavior of prices in the marketplace. As they do, they may change their valuations and switch from being gloomy bears to brave bulls (or the other way around). We have referred to this switching behavior as *adaptive valuations*.

## Three Punch Lines

Our three-part story has moved from fuzzy assessments, to divergent expectations, to adaptive valuations, and it has three punch lines. First, an equity market is a network within which no one participant alone establishes an equilibrium price—rather, the participants collectively search for a price. A property of a network is that the larger it is, the more effectively it operates (we discussed *network externalities* in Chapter 6).

The second punch line of our story is that there is no unique equilibrium price for a stock. That is right, you have read it correctly—a stock does not have a unique equilibrium value. Rather, the price that is ultimately discovered in the marketplace depends on exactly how information about separate valuations is communicated between the members of the group. Who has spoken first, the brave bulls or the gloomy bears? What direction has the sentiment of the group started moving in? Up or down? Emotions may be contagious and self-reinforcing. Early expressions of optimism may result in higher prices prevailing; early expressions of pessimism may result in lower prices prevailing. This is a random equilibria environment. Random equilibria may not be something that you have previously thought about, but it is a topic of interest to economists. We may more commonly believe that, when a market reaches equilibrium, it is a unique, single-valued solution. However, a network can exhibit multiple equilibria and, in our story, an equity market network does.

The third and final punch line is in the spirit of the second. The multiple equilibria environment is *path dependent*. What does this mean? An important overnight news announcement has occurred for a stock, and a new equilibrium value has to be found. Consistent with our story, following the news event, all participants may agree on the direction of the expected price response, but not on the magnitude. So, vis-à-vis each other, some participants will be bullish and others will be bearish. As the bulls and the bears trade with each other, talk with each other, and observe prices in the market, the price in the market evolves between their different valuations. Assuming no further news release, the price will eventually settle on a stable value that we can call an equilibrium. Will this value be closer to the bullish assessment or to the bearish assessment? Path dependency means that the final, stable value that is discovered depends upon the specific path that price has followed in the discovery process. The early shape of the path, in particular, is of critical importance. If the bulls start out being more vocal, the path will start moving up, more participants (but, of course, not necessarily all) will switch to being bulls, and the final equilibrium price will be closer to the bullish valuation. The opposite will occur if the bears speak up first—the path will start moving down, more participants (but, of course, not necessarily all) will switch to being bears, and the final equilibrium price will be closer to the bearish valuation.

## Welcome to Behavioral Economics

The line of thought that we are following opens a door to behavioral economics and, as it does, it raises myriad questions. How do participants behave in this divergent expectations and adaptive valuations environment? How do they communicate with each other? What signals do they respond to? What patterns in the transaction records give participants confidence that a price, right now, for the current, brief trading moment, is reasonable? No trader wants to make a purchase or complete a sale and then, moments later, look stupid because it was a bad trade. Does a pricing pattern reveal that a current price level really is not supportable? If price pierces a support level (the lower bound of a trading range) or a resistance level (the upper bound of a trading range), does this mean that the market will actually be moving lower or higher? It is certainly valid to address empirical questions such as these with the use of statistical analyses of current and past market data.

This is what technical analysis is all about. Technical analysis need not be viewed as a way of predicting values that will be realized further off into the future. Technical analysts commonly describe their work as studying data generated by the actions of people in the market so that they, the analysts, might better determine the right time to buy or to sell a stock. The analysts may not be looking to predict the more distant future, but to assess the color of the market at the current time so as to know just when and how to submit their orders.

In other words, in a nonrandom walk setting, technical analysis can be a valuable tool for assessing price movements that are likely to occur right now, in the current session, as the broad market searches for appropriate consensus values. For instance, a common practice of technical analysts is to locate a support level and a resistance level for a stock, based on the issue's recent trading range. If either of these bounds is neared but not penetrated, a technical analyst might expect the stock's share price to reverse course and stay within its current trading range. However, if either of these bounds is pierced, a signal is sent to the market that the stock's value has indeed shifted to a new level.

Support and resistance bounds, of course, are not readily apparent, but are assessed based on the skill and experience of the trader. The trader might attempt to profit by using a reversal (contrarian) strategy if price appears to be fluctuating between the two bounds, or a momentum strategy if a support or resistance level is pierced.

In the nonfrictionless environment, it is not just the future that is unknown. The current period is also uncertain. Prices do not change simply because information has changed. They change because they are involved in a search process. It is a search for a stable value, for a value that has been validated by the market, for a value that we can call an equilibrium. The search involves volatility. The divergent views on value and the reassessments of individual participants when they learn what others are thinking provide a basis for understanding the magnitude of intraday volatility that we see on a daily basis in the equity markets.

Intraday volatility is high and it matters. When you invest, you might be thinking of holding a position for a year or more but, when you trade, you must be concerned about the next thirty minutes (or seconds). Price movements within the next thirty minutes (or seconds) may have little to do with underlying information change. They may simply reflect the dynamic behavior of the order flow in a nonfrictionless marketplace. You will operate well in this environment not only based on how well you have assessed a stock's fundamentals, but according to your knowledge of its order flow. At this point, you are not trading the stock—you are trading the order flow. When it comes to trading the order flow, you do not research the fundamentals. It is the dynamic properties of the stock's intraday price movements that you want to know about. This is when technical analysis and algo trading are most useful. As you apply it, in whatever way you find has the most value for you given the type of trader you are, we wish you the best of luck for the next thirty minutes (or seconds).

## THE BIG PICTURE ON FINANCIAL MARKET EFFICIENCY

We started this chapter by looking at some fundamental efficiency conditions that theory says would be satisfied in a perfectly competitive, nondynamic, and frictionless trading environment. Then we delved into the intricacies of the nonfrictionless world and considered 10 classic causes of market failure, after which we added one more (noisy price discovery).

Of course, we do not live and operate in a frictionless environment. Quite to the contrary, many markets are not perfectly competitive, the current economic environment is far from static, a spectrum of costs and blockages exists that would not be there to perturb us in a frictionless marketplace, and there are various sources of market failure. We have devoted major attention in this chapter to another way in which the environment is imperfect: we have only limited ability to deal with information sets that are so enormous, complex, incomplete, and imprecise. Nevertheless, the efficient market hypothesis remains elegant, compelling, extraordinarily difficult to refute, and humbling. The EMH stands as a warning to all aspiring young masters of the universe who, armed with what they might think are magic formulas, believe that they will earn millions. Some might, but many, many others more than likely will not.

To repeat, the efficient market hypothesis says that nobody with only publicly available market information and what might appear to be a crystal ball will be able to (1) know the future path that prices will follow and (2) trade profitably on that knowledge. For one thing, prices should, and to a large extent do, reflect existing information. Further, any exploitable predictability in price changes (returns) would be an arbitrage opportunity—shares could be bought at one moment at a relatively low price and sold at a later moment at a higher price (or vice versa with a short sale). As is true with arbitrage, the very act of this trading eliminates an exploitable pattern. And so, EMH devotees think, any and all arbitrage opportunities are exploited and eliminated, the predictable powers of the technical analysts (the Darth Vaders of the trading world) are annihilated, and the EMH (Luke Skywalker) is rescued.

But is technical analysis patently false as our previously cited quote from Burton Malkiel says it is? Thirty-seven years before Burton Malkiel's classic, *A Random Walk Down Wall Street* was published, another classic by Benjamin Graham and David Dodd appeared in the market (both books remain in print to this day). Terms such as "random walk" and "efficient market hypothesis" are not in Graham and Dodd's index, but an interesting quote that conveys their thoughts about these two concepts is worthy of consideration:

> *It follows that the responsibility of managements to act in the interest of their shareholders includes the obligation to prevent—in so far as they are able—the establishment of either absurdly high or unduly low prices for their securities.*
>
> —Benjamin Graham and David Dodd, *Security Analysis* (McGraw-Hill, 1934), p. 515

Two thoughts expressed in the quote are of particular interest: (1) A company's management has some responsibility for maintaining its share price at reasonable levels, and (2) share prices could be "absurdly high or unduly low," which is something that would not happen in a random walk, efficient markets world.

In more recent years, our empirical tests have grown considerably more sophisticated, vast sets of intraday trade and quotations data (which are referred to as *high frequency data*) have become available, and substantial evidence has accumulated that dependency patterns do in fact exist in stock returns. Consequently, the efficient market hypothesis is not looking quite as stellar as it once did. Moreover, one prominent economist, Lawrence Summers, who in November 2008 was appointed Director of the National Economic Council by President Obama, has argued that an inability to reject the EMH is not a sufficient basis for accepting it. Summers writes,

> *The standard theoretical argument for market efficiency is that unless securities are priced efficiently, there will be opportunities to earn excess returns. Speculators will take advantage of these opportunities by arbitraging away any inefficiencies in the pricing of securities. This argument does not explain how speculators become aware of profit opportunities. The same problems of identification described here as confronting financial economists also plague "would be" speculators. If the large persistent valuation errors considered here leave no statistically discernible trace in the historical patterns of returns, it is hard to see how speculators could become aware of them.*
>
> —Lawrence H. Summers, "Does the Stock Market Rationally Reflect Fundamental Values?" *Journal of Finance*, XLI, no. 3 (July 1986), pp. 591–601

Summers subsequently uttered a far stronger statement. Shortly after the dramatic market plunge of October 19, 1987, he was quoted in the *Wall Street Journal* as saying that the Efficient Market Hypothesis is "the most remarkable error in the history of economic theory."[9]

Thirty years earlier, the famous financier Bernard Baruch expressed his view of the link between information and stock prices:

> *The prices of stocks—and commodities and bonds as well—are affected by literally anything and everything that happens in our world, from new inventions and the changing value of the dollar to vagaries of the weather and the threat of war or the prospect of peace. But these happenings do not make themselves felt in Wall Street in an impersonal way, like so many jigglings on a seismograph. What registers in the stock market's fluctuations are not the events themselves but the human reactions to these events, how*

---

[9] *Wall Street Journal* (October 23, 1987).

> *millions of individual men and women feel these happenings may affect the future.*
>
> —Bernard M. Baruch, *My Own Story*
> (New York: Henry Holt, 1957), p. 84

We are playing in a different ballgame if what matters are human reactions to events rather than the events themselves. In this setting, the link between information and security prices is seen to be far more intricate, behavioral considerations attain considerably more importance, human emotions play a significant role, and, in this context, trade and quote records tell a far more interesting story.

## MARKETS AT RISK: THE FINANCIAL TURMOIL OF 2007–2009

We touched on the first part of the history in the preface:

> *The storm clouds had started to gather in the summer of 2007. In early August, the first signs of a severe credit crunch appeared, as several major issuers of asset backed commercial paper, that had provided prior funding of mortgage backed securities, found that they were no longer able to refund their outstanding issues. This evaporation of liquidity spread to other sectors of the capital markets. Over the ensuing year, the illiquidity problem aggravated and, as it did, some major financial institutions such as Bear Stearns weakened. In the fall of 2008, a major financial crisis erupted, and it quickly spread to the broader markets in the U.S. and beyond.*

As summer turned into autumn in the year 2008, micro markets in the United States and around the world were battered by the most powerful economic storm that we have experienced since the crash of 1929. Major institutions tumbled and some disappeared. Bear Stearns is now part of J.P. Morgan Chase. The big insurance company, American International Group (AIG), got on government life support. The venerable investment banking firm of Lehman Brothers went under and is no more. The two giant mortgage companies, Fannie Mae and Freddie Mac, were also put on government life support. Merrill Lynch is now owned by Bank of America, and Wachovia is now owned by Wells Fargo. The last two major U.S. investment banks, Goldman Sachs and Morgan Stanley, are now being regulated by the Federal Reserve as bank holding companies, a change that has subjected them to appreciably tighter regulations and to closer supervision by bank examiners. Citicorp and Bank of America increasingly came under government control.

None of this reconciles very well with a market efficiency story. To many who had not previously given it much thought, it became crystal clear in fall 2008 just how complex the finance function truly is. The financial sector is not simply a conduit for transmitting funds from those who save to those who invest. Financial activity is integrally connected with broader economic activity, it pervades the micro markets. How so? Credit lines were freezing up and nonfinancial firms were heading into bankruptcy. And, as one frightening development lead to another, a further reality was underscored: any major breakdown in the financial sector carries the specter of widespread systemic risk. Reviewing these developments, we are led to conclude that free and reasonably competitive markets can be far from perfectly efficient.

The causes of the turmoil were varied. Extremely low interest rates, easy mortgage terms, and subprime lending spurred a bubble in housing prices. Of course, everything seemed to be okay on the home front as long as housing prices kept rising, but when home values started heading south trouble began to smolder. Fuel was added to the fire as individual mortgages were being bundled into packages and securitized (that is, shares in a package could be bought and sold as a security). With home prices dropping, an alarming percentage of the mortgage-based securitized assets turned toxic. Concurrently, the financial markets experienced an explosive growth in the notional value of new financial instruments such as collateralized debt obligations (CDOs are securitized shares in financial asset pools) and credit default swaps (a CDS is a contract between a buyer and a seller where a buyer makes periodic payments to a seller for the right to receive payment of a sum equal to the principal value of an underlying financial instrument if the underlying instrument goes into default). The markets for these new instruments lacked transparency, they were largely unregulated, and the risks they entailed were seriously misjudged. Include the fact that highly leveraged positions were being attained by a wide spectrum of professional asset managers, and you will no doubt appreciate that the situation had become extremely dangerous.

But that was not all. Underlying the whole complex structure was the unfortunate reality of ignorance. In the opaque environment of the time, the total amount and quality of outstanding positions was not well known or understood. Default probabilities were seriously underestimated, and credit ratings were seriously overstated by the credit agencies. Early liquidations were forced, markets became less liquid, and risk positions ceased to be properly hedged. There was widespread ignorance among the movers and shakers about one further thing: they did not adequately appreciate the extensive interconnectedness of the markets. As the debacle unfolded, a tremendous deleveraging of positions was forced, there were runs on institutions like Bear Stearns and Lehman, and problems spread from financial markets to credit markets, and then to the broader economy. As of this writing, automakers General Motors and Chrysler have danced through bankruptcy and have been rescued by the federal government.

An ant, a little red or black insect, has been classified as one of the dumbest creatures on earth.[10] Yet collectively, ants are very intelligent. Look at how they construct their colonies, divide tasks between themselves, hunt for food and, as a colony, avoid predators. Collectively, ants are very intelligent. So many people in the field of financial economics are highly intelligent: the quants, the financial engineers, the entrepreneurs, the academicians, and so forth. Finance has attracted many brilliant people to its ranks. Yet, collectively, we were not doing so well in the fall of 2008. Collectively, we had run into a startling, frightening hole. Are we exactly the opposite of ants? How can we individually be so brilliant and, at the same time, collectively produce such a dismal outcome?

As noted, the breakdown started in the financial sector with the cheap availability of financing and a plethora of subprime loans for the housing market. From there it spread into the broader market. We talk about Wall Street *versus* Main Street. This is not quite right. Financial markets are absolutely essential for the broad economy. There is a huge connect between Wall Street and Main Street. Financial capital enables firms across the board to operate, just as oil enables physical capital, from bulldozers to automobiles to airplanes, to run. Yet the financial markets are fragile. We do not always think about it and, in "normal" times, we do not even see it. But they are fragile, especially in today's high frequency, electronic environment with large pools of capital that can fly anywhere around the world at a moment's notice.

Volatility and risk are of central importance to those of us in financial economics. They lie at the center of virtually all of our work. Without volatility, without risk, finance would not exist as a subject in our business school curriculum. Finance, quite simply, would not be distinguishable from deterministic economics. Neither would finance departments in banks and industrial firms be endowed with anywhere near the importance that they currently have.

*Risk aversion*, *risk hedging*, *risk management*, *value at risk*, *risk measurement*, and *risk premium*: terms like these are part of everyday parlance in the world of finance. In the industry we have high-powered minds, high-powered valuation formulas, high-powered trading algorithms, and high-powered electronic technology to pull it all together. And yet, the events of 2008 have shown us what risk really is, and how at risk the financial markets truly are. Now our buzz words also include terms such as *subprime mortgages*, *government bailout*, *moral hazard*, *toxic assets*, *too big to fail*, and *systemic risk*.

---

[10] Material on this part of the section draws from Robert A. Schwartz, "Markets at Risk," *Journal of Trading* (Spring 2009), 46–49.

Unfortunately, there is much about volatility and risk that we do not understand. It appears that some of the most important decision makers in the hallowed halls of Wall Street had lost sight of the fact that risk isn't the only contributor to volatility. Risk has a well-defined meaning for economists, and we devoted considerable attention to it in Chapter 3. As you know, risk exists when an outcome can be described as a draw from a probability distribution that has known parameters. Flip a fair coin and bet on the outcome: the chance of heads equals 50 percent, and the chance of tails equals 50 percent, but beyond that we do not know what the outcome will be until after we have flipped the coin. That is risk. Give us the probability distributions and we will do a good job modeling it.

But along with risk, there is also *uncertainty*. With uncertainty, we do not know the probability distributions, or we might think we know the probabilities but have misstated them and, at times, we may not even know what all of the possible outcomes are. Dealing wisely with uncertainty is a huge challenge. Insufficient formal attention has been given to uncertainty as a cause of volatility.

Risk and uncertainty coexist with dynamic change, an integral part of a modern, free market economy. We value change because of the enhanced array of goods and services, and the generally higher standards of living, that economic growth provides. But innovation also has its costs, and these are indubitably borne disproportionately by different households and firms. Adam Smith's invisible hand can offer both carrots and sticks. There is an irrevocable interplay between growth and destruction, which we see in Schumpeter's concept of creative destruction.[11] Creative destruction is new products driving out old products (for example, compact discs are being replaced by MP3 players); it is the introduction of new skills (such as those based on technology and automation) that make older skills obsolete. As this occurs, some older firms are driven out of business, some untrained workers get laid off, and the "destruction" accelerates in times of recession. What does this say about the efficiency of a free market from a social welfare perspective? It sure gives us a lot to think about (and to debate about).

---

[11] The economist Joseph A. Schumpeter introduced the concept of creative destruction in his 1942 book, *Capitalism, Socialism, and Democracy,* reissued by Routledge, London, 1992.

Earlier in the chapter we enumerated ten classic causes of market failure. Let's revisit them and get a taste of the role they each played in the debacle of 2008.

1. **Imperfect and asymmetric information:** In fall 2008, the markets were fraught with poor information. People had only very imperfect knowledge of, for instance, the total value of toxic assets outstanding, or of the total notional value of instruments such as CDSs, which were written and traded in markets that lacked adequate transparency. When information is imperfect, some people are undoubtedly better informed than others and, at the time, the asymmetry of information possession further impeded the structuring of mutually beneficial deals.

   A particularly stark example of informational asymmetry concerned the collapse of Lehman. Potential buyers of the failing enterprise, being at a decided information disadvantage vis-à-vis Lehman insiders concerning the value of Lehman's assets, shied away from making a deal that could have rescued the company.
2. **Irrational behavior:** The highly volatile markets that followed the financial collapse provide a good example of how emotions can affect decision making. We saw in Chapter 2 how a rational investor will select a portfolio that offers an optimal balance between risk and expected return. Returns are, of course, good—we all strive for them—while risk is undesirable and most of us require adequate compensation before accepting it. But think of a participant who, armed with a string of stunning successes, puts too little emphasis on the true riskiness of a situation. How would you describe this individual who, in a foolhardy effort to nail down yet another glorious win, recklessly discounts risk? "Greedy" would be an appropriate response. Leading up to the debacle, any number of investment bankers (and others) seemingly became greedy.

   Now think of a participant who, stung by a string of big losses, shifts focus, deeply discounts expected returns, throws in the towel, and capitulates. How would you describe this person's emotions? "Fear" is the appropriate concept. Following the onset of the widening financial trouble, fear spread and the markets kept on tumbling, faster and further.

   These two emotions, "greed" and "fear," got a lot of play in the media in fall 2008. As they impact investor decisions, prices go up too high in a rising market and drop too low in a falling market. A term that we have previously introduced, *adaptive valuations*, interacts with this behavior. As greed and fear spread, *overshooting* is the term we use to describe the price patterns that follow.

3. **Short-run myopia:** Top management in the financial firms was under strong and continual pressure to show sparkling bottom-line results (as is the case for many firms). In the years leading up to the debacle, some hefty profits were being realized by some investment banking firms that were taking highly leveraged risk positions, and their shares were trading at lofty prices. A firm that did not participate in these profit opportunities because of a more cautious assessment of risk did not look very good in the short run, but would have been proven wise in the long run. However, the pressure to join the party was great, and short-run thinking crowded out longer-run judgment.
4. **Externalities:** The crisis was replete with externalities. The demise of Lehman weakened other firms, and fear spread that it would lead to further bankruptcies. The collapse of Fannie Mae, or Freddie Mac, or the giant insurance company American International Group (AIG) would have had widespread repercussions. So would the collapse of Merrill and Goldman Sachs. AIG, for instance, was a major trading partner of Goldman, and when the mortgage crisis pushed the big insurance firm to the brink of collapse, frightened investors and clients pulled out of Goldman, and Goldman shares plunged.

   Fearing widespread fallout, bank lending shut down and the channels of credit froze up. These spillover effects are externalities. The term "too big to fail," which a decade earlier had been applied to the giant hedge fund Greenwich Long Term Capital, captures the thought: Huge firms have extensive impacts on other markets and, if a firm is large enough, the widespread damage it could create by collapsing is so threatening to the broader market that other firms and/or government intercede to prevent its demise.

   "Too big to fail" specifically means that the negative externalities that would attend failure are too big to let the failure occur. These externalities underlie systematic risk, which we consider in point 10 below.
5. **Network externalities:** An equity market is a network that exhibits network externalities. Participants, by placing their orders in the market, provide liquidity to each other, and they provide more liquidity to each other for larger capitalization stocks, such as ExxonMobil and Wal-Mart (which have more shareholders) than they do for mid- and small-cap stocks. But for the traders individually, the liquidity benefits that they bring to the network is an externality, and thus aggregate liquidity is undersupplied by the free market.

In the turbulent months of 2008, liquidity dried up throughout much of the financial system and government was called on as a liquidity provider of last resort. These events have clearly demonstrated the need for a facilitator in times of crisis. However, even under less stressful conditions, liquidity creation does not just happen automatically; it often calls for a facilitator (although not necessarily for government per se) to play this role.

6. **Moral hazard:** Moral hazard situations can exist in many different contexts. An interesting example for the potential of a moral hazard situation involves the use of credit default swaps. A CDS resembles an insurance contract, but has not been regulated as such. With an insurance contract, it is required that the individual buying the policy have a personal interest in the item or person being insured. For instance, to buy life insurance on somebody, the person whose life is being insured must be a family member, a business associate, or someone else of clearly identifiable importance to the person who is buying the insurance. At the time, this was not required for a CDS: the contract could be between two people without either of them having an ownership position in the underlying asset. A moral hazard problem could arise if the person buying the CDS (the "insured"), being in a position to profit from the demise of the underlying asset, actually has the capacity to increase the chances of default.

   Government bailout programs can lead to moral hazard. It is one thing to save a failing firm today. It is another thing for firms in the future to believe that they too will be bailed out if they get into trouble. If a risk pays off and a high return is achieved, the firm keeps the winnings; if the firm realizes a cripplingly low return, it is bailed out. Providing such a safety net to an individual firm encourages excessive risk taking, and that is a moral hazard.
7. **Principal-agent problems:** Principal-agent problems are always with us, and they were certainly striking during the market turmoil. As we have said, contracts are incomplete, monitoring is imperfect, and this can lead to excessive compensation and actions on the part of top management that are not in the best interest of the shareholders. High salaries, outsized bonuses, and golden parachutes were all too apparent during the debacle of 2008–2009.
8. **Public goods:** A primary function of a securities market is to establish the prices of the traded assets. We have called this *price discovery*. It is not a simple process. As we have noted, textbook-efficient, equilibrium values are continuously being sought but rarely are they achieved. Prices set in a marketplace are obviously of prime importance to the participants in the trades, but they also very much matter to a broad

spectrum of other participants. Portfolios are assessed using market prices, margin requirements are based on market prices, derivative prices depend on the prices of the underlying assets, and so on and so forth.

A price that is set in a marketplace resembles the beam from a lighthouse that any ship passing in the night can see. A market-determined price, like the beam from a lighthouse, is an informational signal. Collectively, we benefit from more accurate price discovery. Achieving this calls for markets that are appropriately transparent, sufficiently liquid, and not excessively volatile. Unfortunately, these properties of the financial markets suffered badly in 2008–2009. For the most part, the equity markets were relatively transparent, but the markets for other financial products were not, liquidity in many markets dried up, and volatility went through the roof. The public goods properties of the market suffered; quite clearly, better market structure was needed.

9. **Technological inertia:** It is not easy to change the structure of a market. People are accustomed to the old ways of doing things. Firms have found their competitive niches within a market structure and do not want to see them disappear because of a market structure change (think again about Schumpeter's concept of creative destruction). Moreover, new ways of doing something are often not well understood, and plain old ignorance all too often can be a major cause of technological inertia.

   The collapse of the housing market along with a fistful of major financial firms in fall 2008 sent waves of volatility through the financial markets. Without question, share values were bound to plunge in the face of this massive, negative information event. But falling prices is one thing—volatility is another. The volatility encompassed plunging prices that at times were interspersed with soaring prices. The crucial question to be addressed is whether or not the volatility was excessive. The million (or possibly trillion) -dollar question was, would the huge price swings have been more effectively contained if the securities markets had been better structured?

   In recent years, exchanges around the world have considered various ways to enhance the quality of their markets. Nevertheless, progress takes time and technological inertia remains an impediment to reckon with.

10. **Systemic risk:** The collapse in 2008 included a systemic breakdown that encompassed countries around the world. Its origins were in the housing sector. Because widely held financial assets such as mortgage backed securities and their derivatives are tied to the housing market, the crisis quickly spread from the housing sector to major banks and brokerage

houses. The venerable firm of Lehman Brothers went out of business, and the giant insurance company AIG nearly collapsed. But finance pervades the economy, and thus the problems were transmitted to companies in other industries. Most strikingly, the three big American auto manufacturers (General Motors, Ford, and Chrysler), which were already in a weak state, were pushed toward bankruptcy. As the fallout spread, the U.S. economy was heading straight into a recession.

## THE GREATER GENERALITY OF THE FRICTIONS PROBLEM

Our discussion of the effect of market frictions has paid particular attention to one specific marketplace: an equity market. But transaction costs exist in all markets, demand and supply conditions in all markets are unstable over time, and all markets require some mechanism for finding the prices at which trades are made. Could it be that the problems facing the equity markets are symptomatic of difficulties that are more pervasive throughout the economy?

Conditions change rapidly in the equity markets, transaction costs impede trading, and price determination is a complex, imperfect process. Conditions may not change nearly as fast in other markets, but if trading frictions and blockages are substantial, price discovery in other markets may be impeded, and the prices of other goods and services may fail to achieve desired equilibrium values even if the markets are reasonably competitive. Consider, for example, labor markets where long-term contracts are negotiated.

A labor contract, like all contractual agreements, is not easily changed. Because it specifies and fixes the wage rate (along with other terms of employment), it prevents wages from responding freely to the changing forces of supply and demand. As such, we can consider a labor contract a market friction. One may question, therefore, how participants in labor markets respond to the friction, and how their responses affect wage determination and employment. Assume a large number of firms (not necessarily in the same industry) and a large (but not necessarily homogeneous) labor force. Let each firm sign a multiyear contract with its employees at a wage rate that is related to the general level of wages. Then consider the following:

1. What is the actual process by which the general level of wages is determined?
2. Do transaction costs and imperfect information impede the disclosure of demand and supply functions in the labor market as they do in the stock market?

3. Does uncertainty exist concerning the level of wages that would clear a market in the current period and, if so, how do firms and workers respond to this uncertainty?
4. What are the institutional arrangements for dealing with price discovery in labor markets? Staggering contracts would be one such arrangement. (Contracts are staggered when they do not expire at the same time for a set of firms or industries.) Does staggered contracting in labor markets facilitate tracking an equilibrium wage, much as continuous trading may facilitate tracking equilibrium prices in the equity markets?
5. To what extent is unemployment in labor markets a result of contracting, wage uncertainty (which resembles the transaction price uncertainty that we discussed earlier in this chapter), and the labor market in aggregate making mistakes?
6. To what extent might change in the magnitude of wage uncertainty, which may be caused in part by changes in aggregate economic policy, cause fluctuations in employment, unemployment, and output?

Dynamism and allocation efficiency are two powerfully positive attributes of a free market. On the other hand, for a free market, instability is a serious Achilles heel. The occurrence of periodic recessions has long characterized free market economies, and the world endured the start of a frightening economic collapse in 2008. In the fourth quarter of 2008, micro markets around the globe were hit by tidal waves of volatility. Fingers were pointed at many things including the housing bubble, greed, and various technical factors ranging from some accounting rule changes to the absence of certain short-selling restrictions. Fingers were pointed at management failure, government failure (from an operational perspective), regulatory structure failure, and market structure failure. At this point, we call your attention to regulatory intervention and market structure in particular. These two things, if properly designed and implemented, could do much to better stabilize our markets in a risky and uncertain world. Is superior market structure and regulatory design possible? Of course, better is always possible.

It is not a matter of free markets versus regulated markets, or of deregulation versus reregulation. Regulation is indeed needed, but it must be appropriate. The issues, the concerns, the market failure realities upon which regulations should be based must be better understood. The realities of government failure also have to be taken fully into account. Excessive regulation and/or ill-structured regulation can be extremely costly to financial markets in particular and to all markets in general. We turn to issues concerning government regulation in the next chapter. In the meantime, one thing is for sure: The market turbulence of 2008 and beyond has given us a great deal to think about concerning the efficiency of the micro markets.

## CHAPTER SUMMARY

Microeconomic theory presents an intriguing and deeply insightful depiction of how households and firms make optimal consumption and production decisions so as to maximize their feelings of well-being (which we call utility) and their profits. Microeconomic theory also takes a second major step: it examines how the individual, self-serving decisions of households and firms interact on the aggregate level. Aggregation is not a simple matter. Until a careful assessment is made, one cannot know if what is good for participants individually is good for everybody collectively. In this chapter, which has market efficiency as its topic, we have not focused per se on the efficiency with which actual households and firms make their decisions, but on the quality of outcomes that the micro markets can deliver for the broad set of participants in aggregate. Here are the key points that we have made:

- First the good news. In a plain-vanilla, deterministic (nonstochastic), frictionless environment where all participants (firms and households alike) are atomistic, the equilibrium conditions have excellent properties from a social policy point of view. For all goods, the marginal rate of substitution in consumption is the same for all consumers, the marginal rate of technical substitution in production is the same for all firms, and all prices (which reflect utility on the margin) equal their respective marginal costs of production. With the marginal conditions satisfied, no mutually utility-enhancing (for households) or profit-enhancing (for firms) redistribution of resources between economic agents is possible.
- Real-world markets, however, are not ideal environments, and they can fail to deliver even close to textbook-perfect results. In some extreme cases, a micro market may even fail to operate at all (that is, a market might totally collapse). We considered 10 classic causes of market failure: imperfect and asymmetric information, irrational behavior, short-run myopia, externalities, network externalities, moral hazard, principal-agent problems, public goods attributes, technological inertia, and systemic risk.
- A daunting challenge facing the micro markets is the complexity, the size, the incompleteness, and the imprecision of much of the information that households and firms base their consumption and production decisions on. With an eye towards the equity markets, we differentiated between *market information* (for instance, quotes, transaction prices,

and trading volume) and *fundamental information* (for instance, financial information, management strategy, economic information, structural change, and organizational efficiency). We also subdivided information into *public information* (which is widely disseminated and readily available), *inside information* (possessed by only a select set of insiders), and *private information* (generated by individuals through their own analyses of public information).

- Participants base their expectations about future outcomes on the information that they currently possess. For a set of participants, expectations can be *homogeneous* (they all form identical expectations) or *divergent* (while having access to the same public information, different people do not interpret it the same way and thus come to different conclusions about its meaning). Important formulations (such as the Capital Asset Pricing Model) are based on the assumption that expectations are homogeneous. In real-world markets, however, divergent expectations is the name of the game, and this paradigm shift has major implications (concerning, for instance, how equilibrium prices are discovered in a micro market).
- Informational efficiency is a multidimensional concept. We considered the efficiency with which investors make decisions with respect to existing information, the efficiency with which they pursue information-gathering activities, the informational accuracy of equilibrium prices, and the dynamic efficiency of information dissemination. These are not simple matters. A lot is required for a market to be informationally efficient.
- As we turned next to a test of market efficiency, our attention remained focused on equity markets and on informational efficiency. How does one know, you might wonder, whether or not all existing information is fully reflected in stock prices? It turns out that there is a very clever way to infer whether or not prices are being efficiently set. The way to do so involves the *Efficient Market Hypothesis* (EMH). The hypothesis is that information is fully reflected in stock prices, and the best-known test of the EMH is whether or not stock prices follow random walks.
- Here is the logic behind the random walk tests. If all information is reflected in a stock's price, then there is no way to predict how the stock's price will change in the future, because we would have no way of knowing what the new information will be. But what if future stock price changes are correlated with past stock price changes? With a correlation pattern, future price changes could be predicted from past price changes. But predictability itself is information, and acting on this information eliminates the very correlation patterns that it is based on. Once any correlation message is incorporated in price, the correlation patterns can no longer exist and price must be following a random walk.

- Earlier tests broadly supported the random walk hypothesis. But with the advent of high-frequency (intraday) data and new econometric techniques, evidence is accumulating that correlation patterns do exist in stock returns. The random walk hypothesis is not as stellar today as it once was.
- Having established that correlation patterns do exist in stock returns data, we introduced a fifth informational efficiency criterion: the efficiency with which prices are discovered in the marketplace. Having previously presented ten classic causes of market failure, we added noisy price discovery as an eleventh cause.

  Price discovery is not fully efficient to the extent that it works its way out over time while trades are being made. The market is always searching for a sustainable equilibrium value but does not necessarily find one before a demand shift changes the equilibrium value that the market has been trying to discover.
- We clarified that the efficiency of price discovery depends in part on the structure of the marketplace where the trades are being made. The complexity of price discovery also reflects the fact that investors are less than textbook efficient information processing machines.
- Our price discovery story has three parts that move from fuzzy assessments, to divergent expectation, to adaptive valuations. And it has three punch lines: (1) A market is a network within which no one participant alone establishes an equilibrium price, (2) a stock does not have a *unique* equilibrium price (its price, which is discovered in the marketplace, depends on how the separate valuations of the various participants are communicated between the members of the group), and (3) the equilibrium value that is discovered for a stock is path dependent. Namely, after any shift in investor desires to hold shares of a stock, the path that price then follows during the early stages of the adjustment process affects the new value that the path will eventually converge on.
- Recognizing price discovery as a dynamic process accomplishes several things. It opens the door to behavioral economics; it provides a raison d'être for technical analysis and algorithmic trading; it offers an explanation of why intraday prices are as volatile as we observe them to be, especially at the opening of a trading session and then again at the close; and it underscores the reality that market structure matters, that the efficiency of market outcomes depends in part at least on the quality of a market's structure.
- The big picture that we have presented on financial market efficiency is that markets may be a good deal less than perfectly efficient. For many practitioners, this will not appear to be a revelation. But microeconomists have long seen and respected the inherent efficiency of free and

competitive markets, and financial economists have found additional validation in their empirical studies of stock price movements. We respect these findings. The point is, the equity markets are not 100 percent efficient. Should we look at the percent of the glass that is full, whatever that percentage might be (say 90 percent), or at the percent that is empty (say 10 percent)? We call attention to "the 10 percent," and we stress the importance of seeing it as such. Our quotes from Graham and Dodd, from Lawrence Summers, and from Bernard Baruch are in keeping with this thinking.

- The financial turmoil that started in the fall of 2008 and hit markets around the world underscored the problems that can beset the micro markets. We turned to the recent (and ongoing) events in the next section of the chapter. In so doing, we first paid attention to four concepts: risk, uncertainty (which is differentiated from risk), volatility (which can characterize free and competitive markets), and systemic risk. We then revisited the ten causes of market failure to get a taste of the role that each of them played in the debacle of 2008.
- The final topic in this chapter was the greater generality of the frictions problem. We have focused on one market: the equity market. But to better understand one market is to open one's eyes to realities that may exist far more broadly in the micro markets. Friction (including imperfect information) clearly impacts the workings of the equity markets. How about other markets? The equity markets move with lightning speed—as of 9:30 in the morning, 12:00 noon might be the "long run." Housing markets, labor markets, and most other markets move at a far slower pace. Nevertheless, they all face the frictions problem. Is price discovery unique to the equity markets? The very writing of a contract, once it has been signed, imposes an adjustment cost or, in our terminology, a friction. How are prices determined while a contract is being written? Matters such as the negotiation of an administered price aside, price discovery is certainly an issue in, say, the writing of a three-year labor contract. We underscore the importance of understanding the issue.

We ended the chapter with the thought that it is not a matter of free markets versus regulated markets, or of deregulation versus reregulation. Appropriate regulation is certainly needed. But how do we get it? What is the proper balance between the discipline inherent in a free market and the rules, regulations, and enforcements imposed by government? In this chapter, we focused on the abilities of free markets to deliver desired results, but we have also devoted much attention to the inefficiencies of a free market. Does this call for government to step up and play a more decisive role? In the next chapter we turn to issues concerning government regulation.

## LET'S CONSIDER AN APPLICATION: HOW ARE EQUILIBRIUM PRICES DISCOVERED IN A COMPETITIVE MARKET?

Solving for an equilibrium price using an economic model of a competitive market is one thing; actually discovering an equilibrium value in a real-world, nonfrictionless environment is another matter. In the highly stylized competitive market, the intersection of a demand curve and a supply curve identifies the equilibrium price. In an actual market, transaction prices are set by the interaction of buy and sell orders, and while economic forces might push transaction prices toward equilibrium values, there is no guarantee that equilibrium will be attained quickly and then closely adhered to. After all, no one really sees the demand and supply curves, and no one actually knows exactly what the equilibrium value is. The problem of finding equilibrium values is apparent in the equity markets, and it is to these markets that we once again turn to gain insight into this important economic activity.

This chapter has presented price discovery as a critically important economic function of an equity market, In this application we devote further thought to the processes by which prices are set in an actual equities market.* We structure the analysis as follows:

- Assume a large-cap, highly liquid stock that has a sizable number of shareholders.
- Assume that participants' orders to buy and to sell shares of that stock are all sent to one equity market, "the exchange."
- Let each participant trade just one round lot of the stock, as either a buyer or a seller.
- Give some consideration to an environment where analyst expectations are homogeneous, but for the most part assume that analyst expectations are divergent.
- Give some consideration to call auction trading on the exchange, but for the most part focus on a continuous order-driven market that employs a transparent limit order book as its trading platform.

---

*Material in this application section draws from Puneet Handa, Robert A. Schwartz, and Ashish Tiwari, "Quote Setting and Price Formation in an Order Driven Market," *Journal of Financial Markets*, 6 (2003), 461–489; and from Jacob Paroush, Robert A. Schwarz, and Avner Wolf, "The Dynamic Process of Price Discovery in an Equity Market," *Managerial Finance*, 2010, forthcoming.

(*Continued*)

(*Continued*)

- In the continuous order–driven market, participants arrive one at a time in random sequence. Upon arrival at the market, a participant observes the limit order book and either submits a market order or places his or her own limit order on the book.
- Assume that no informational change occurs during the trading interval. All that happens is that buyers and sellers arrive at the market in random sequence and trade with each other because they have different (divergent) valuations for the stock.

Here are our questions:

1. How applicable do you believe the perfectly competitive model is to the market that we are analyzing: a highly liquid, large-cap stock that is traded only on one exchange?
2. Where do you believe price discovery would take place for the large-cap stock if all relevant information concerning the stock is publicly known, and if all informed participants interpret the information in the same way (that is, form homogeneous expectations): (1) in the upstairs offices of the research analysts, or (2) in the stock exchange where customer buy and sell orders meet and are turned into trades?
3. As a follow-up to question 2, where do you believe price discovery would take place if informed participants do not all interpret information identically (that is, have divergent expectations)?
4. We can represent divergent expectations by assuming just two valuations. With reference to the dividend discount model that we used in the chapter to show the sensitivity of stock price to growth rate expectations, assume that the relatively bullish participants are expecting an annual growth rate of 7.545 percent and are valuing the stock at $55, while the relatively bearish participants are expecting an annual growth rate of 7.000 percent and are valuing the stock at $45. Are you comfortable with this simple representation of divergent expectations?
5. Regarding the two prices mentioned in point 4 ($55 and $45): Are they equilibrium values or reservation prices?
6. Assume *N* participants, *k* percent of whom are bullish (and 1− *k* percent of whom are bearish). Let *k* = .75. Assume the exchange offers call auction trading only, and that the participants' orders

reflect their share valuations ($55 and $45). Should an investor submit a buy order at $55 if that is his or her assessment (or a sell order at $45 if that is his or her assessment) when sending an order to a call auction? To answer this one, you might want to flip back to Chapter 4 and the discussion surrounding Exhibit 4.8.

7. As a follow-up to question 6, plot the aggregate market buy and sell curves that have been submitted to the call. Is the equilibrium price uniquely determined? If not, how do you suggest that the market clearing price be set?
8. Next, switch gears and consider a continuous order–driven market. We dealt with this case in Chapter 4 and can pick up here where we left off there. In brief, the optimal bid prices that will be set by the bullish participants and the optimal ask prices that will be set by the bearish participants can be solved for if $k$ (the percentage that is bullish) is known. If $k = 0.5$, the quotes will straddle $50, the average of the two valuations. If $k > 0.5$, the quotes will be closer to $55, the higher valuation. If $k < 0.5$, the quotes will be closer to $45, the lower valuation. Does this make intuitive sense to you?
9. Obtaining the equilibrium noted in point 8 involves solving a set of simultaneous equations. In doing this, we assume not only that all participants know $k$, but that they all also know how everyone else will act, and that they all know that everybody knows all of this (the so-called "common knowledge" assumption). Consequently, all buyers come up with the same optimal solution for the price of their buy limit orders and whether to instead place a market buy order. Similarly, all sellers come up with the same optimal solution for the price of their sell limit orders and whether to instead place a market sell order. In this setting, will there be any reason for the quotes to change over the course of the trading period as the buyers and sellers arrive in random sequence?
10. With the continuous order driven market, for any given set of valuations (such as $55 and $45), in which of the following three cases is the divergence of expectations the greatest: $k = 0.50$, $k = 0.75$, or $k = 1.00$ (or, equivalently, $k = 0.50$, $k = 0.25$, or $k = 0.00$)?
11. With the continuous order driven market, how will the market clear? This is a tricky one. Perhaps $k = .75$. Then there are $.75N$ participants seeking to buy one round lot, and $.25N$

(*Continued*)

(*Continued*)

participantsseeking to sell one round lot. Even if all of the bearish participants succeed in selling their one round lot, .50*N* buy orders will not be filled. And yet, given the quotes in the marketplace (which, we remind you, are strictly below the $55 that the buyers would just be willing to pay if they had to), no explicit rationing mechanism is needed and no buyer will feel slighted. How come? What enables the market to clear? (Hint: Think of the willingness of a limit order trader to accept the risk of not executing for the possibility of achieving an execution at a better price.) As a follow-up, under which market structure do you expect transactions volume to be greater: the continuous market or the call auction?

**12.** And now for three big realities: (1) *k* is not observable before the start of a trading session, (2) participants gain knowledge of *k* as they see buy and sell orders arriving at the market, and (3) participants can change their valuations while trading progresses and they learn the valuations of others. That is, if the value of *k* is revealed to be relatively high, some participants will change their valuations from $45 to $55, and if the value of *k* is revealed to be relatively low, some other participants will change their valuations from $55 to $45. These three realities give rise to the following questions:

- What does this adaptive valuation behavior imply about the path that prices might follow in the process of attaining a stable equilibrium value?
- Do you see why the current formulation implies that the value a price can converge on is path dependent?
- What does the path-dependent process imply about momentum trading?
- The application at the end of Chapter 3 dealt with bandwagon and snob effects. How does this relate to the current formulation?
- Do you sense intuitively that the process that we have described can explain elevated levels of intraday volatility, especially around market openings?

## TraderEx SIMULATION: FINDING AN EQUILIBRIUM VALUE

Again we turn to the TraderEx simulation software that is available in *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course*. Finding equilibrium values is an important function for any marketplace. Our application in this chapter has dealt with price discovery in an equity market. The simulation exercise for this chapter stays with this. The objective of the exercise is to contrast the process of price discovery under two different environments (A and B):

**A.** The underlying equilibrium price, because it is not sensitive to the incoming order flow, is relatively stable.

**B.** The underlying equilibrium price, because it is sensitive to the incoming order flow, is relatively unstable.

Here is the setting:

- You submit your limit and market orders to a transparent, plain-vanilla limit-order book.
- Each simulation run will be for one trading day (roughly 10 minutes of clock time).
- Run several simulations in each of the two environments as a day trader. Towards the end of each run unwind your inventory position and close the trading day with zero shares (or as close to zero as you can get).
- Your objective can be to maximize your trading profits, but achieving this goal is not the reason for running the exercise. Rather, pay attention to the prices that are set in the simulation run during the trading day. To this end, record the mid-spread of the bid-ask quotes for each hour on the hour, from 10:00 A.M. through 4:00 P.M.
- Take your recorded mid-spread prices for each of the runs and, for each of the two environments, assess their distributions for each of the seven hours of the trading day.
- Additional details for the simulation exercise are provided in the TraderEx book.

Please respond to the following:

1. For each of the two environments, A and B, contrast the dispersion of the mid-spread prices across your simulation runs for each of the seven different hours of the trading day. Do you see any pattern?
2. Contrast the dispersion across your simulation runs for the A versus B environments as of 10:00 A.M., 12:00 noon, and 4:00 P.M. For which

(*Continued*)

(*Continued*)

environment is the dispersion greater? If you do not observe a clear difference, run the simulation a few more times to increase the size of your sample.

3. The two simulated environments have been structured so that the variability of prices from one run to another, at any specific hour of the trading day, is expected to be greater in environment B than in environment A. This is because the equilibrium price is sensitive to the order flow in environment B while it is not sensitive in environment A. Question: what makes the equilibrium sensitive to the order flow? Answer: (1) the computer-generated order flow reflects the divergence of expectations, (2) the proportion ($k$) of investors who are bullish is revealed as trading progresses, and (3) the computer-generated order flow reflects the adaptive valuation behavior of participants. Can you see the evidence of this in the data that you have collected?
4. In some of your simulation runs, acquire a large long position by bunching your buy orders in the morning, and then work the inventory position off by bunching your sell orders in the afternoon. Did this bunching action affect the closing prices and, if so, in which environment did it have a bigger effect, A or B? Explain your findings.
5. Explain what the simulation exercise has demonstrated concerning momentum moves, correlation patterns in returns, and the short-term volatility of prices.
6. How would you rate the efficiency of price discovery in the two different simulation environments?

# CHAPTER 8

# Public Policy and the Interplay between Competition, Technology, and Regulation

Micro markets around the world are being shaped by three powerful forces: competition (both domestic and global), technology development, and regulation. Each of these drivers may be viewed as desirable or well intentioned, but their joint impact is producing results that are difficult to predict, hard to control, and not easy to understand.

Competition, which we focused on in Chapter 6, is very much a product of technology and regulation. Technology opens up new markets and new ways of operating, while regulation establishes the rules by which firms in an industry can compete. Reciprocally, technology development is itself in large part attributable to competition as firms innovate to boost their revenues, cut their costs, and generally strengthen their positions in the marketplace. As technology and competition evolve in the micro markets, a spectrum of public policy issues enters the picture. In considering them and the regulatory issues involved, we continue to pay particular attention to one specific micro market—the equity market.

As of the mid-1970s, equity markets around the globe had experienced little structural change for the previous two centuries. Since then, the markets have been transformed by massive technology development and intense intermarket competition. Concurrently, major public policy issues have arisen that have been debated by government officials, industry participants, and academicians, and major regulatory initiatives have been taken.

Public policy issues commonly lead to government regulatory intervention in the workings of a micro market. Alternatively, an industry can regulate itself. Self-regulation is a branding feature of any membership organization (for example, golf clubs and dining clubs typically impose rules of behavior and dress codes for their members). In the United States, stock exchanges such as the NYSE and NASDAQ have formal *Self-Regulatory Organization* (SRO) obligations. We discuss their SRO role in a later section of this chapter.

Public policy as it pertains to the regulation of a micro market has four major targets:

1. **Honesty:** Rules are specified and enforced to prevent criminal behavior; for the equity markets, this includes the manipulation of prices, the exploitation of inside information, and various other abuses of power and position.
2. **Free-market outcomes:** Prices and quantities set in an unregulated market may be deemed undesirable from a public policy perspective and, accordingly, may be altered by government intervention. Examples include minimum wage legislation, agricultural price supports, rent control, other price controls, support of education, restrictions on fuel emissions, sales taxes, subsidies, and so on. We considered some of these interventions in Chapter 3. In addition, antitrust legislation can call for the breakup of some monopoly firms, natural monopolies may be price-regulated, and certain forms of price discrimination are disallowed. We considered these issues in Chapter 6.
3. **Macroeconomic activity:** Monetary and fiscal policy are the primary macroeconomic tools that are employed to dampen inflation, to support employment, in general to dampen business cycles, to encourage economic growth, and to rebalance the distributions of income and of opportunities across participants.
4. **Market structure:** The very structure of a marketplace can be subjected to government regulatory control. In light of their importance to the broader economy, the structures of the financial markets in particular have been so subjected. A financial market's architecture is for the most part defined by its rule book—that is, the rules and regulations (regs) that determine how orders are submitted, handled, and translated into trades and transaction prices. This chapter focuses on this fourth target, paying particular attention to government involvement in the structure of the equity markets.

Much oversight and regulation can and does come from the private sector and, accordingly, the first section of this chapter considers the governance structure and self-regulatory obligations of a securities exchange. In the second section, we turn to technology developments (including automation) and, in so doing, pay particular attention to the dramatic impact that technology has had on the structure and operations of one specific organization, the NASDAQ Stock Market. The remainder (and bulk of) the chapter then deals with government regulation.

With regard to government regulation, the financial markets are our specific focus in the remaining sections of the chapter. After first reviewing the government regulatory focus as it has evolved since the 1930s to include honesty, then the efficiency of the market for broker/dealer services, and then the efficiency with which equity shares are traded and priced, we consider various reasons why government regulatory intervention may itself be flawed; this discussion parallels our enumeration in Chapter 7 of reasons why the free market may fail to deliver fully efficient results. A substantial portion of the chapter is then devoted to seven specific equity market initiatives that have been taken by the U.S. regulatory authorities since 1975. Each of these initiatives has had major economic consequences, and economic considerations have permeated the discussions and often heated debates concerning their adaption.

As you follow the debates, you will see microeconomic theory in action. Were the interventions successful? Along with expressing some of our own thoughts, we encourage you to come to your own conclusions. That said, for those whose bent is toward greater government intervention in the workings of the micro markets, we conclude the chapter with seven caveats.

## GOVERNANCE STRUCTURE AND SELF-REGULATION

A securities exchange may be a for-profit (privatized) organization that is run for the benefit of its shareholders, or it may be a non profit, mutual organization that is run for the benefit of its members. An equity-based, profit-maximizing firm has one overriding objective: maximize the price of its equity shares. A membership organization has a different overriding objective: maximize the value of access privileges, which is reflected in the price of membership. With regard to an equities market, the investors who come to the market to trade are the customers of a privatized (for profit) exchange. In contrast, the broker/dealer members, not the investors, are the customers of the membership organization. Put it all together, and the difference in organizational structure has major implications for exchange operations and regulation.

### Demutualization

Exchanges have historically been mutual organizations. In the United States, an exchange is defined as a meeting place for intermediaries. In the precomputer era, intermediaries were needed (1) to handle the trades (as dealers or as brokers), and (2) to control *counterparty risk* by knowing and standing behind their customers (if a customer fails to deliver either cash or shares, the customer's broker is obligated to make the delivery). On both counts, computer technology has paved the way for demutualization: (1) It has enabled disintermediation (orders can be entered and executed electronically), and (2) it has made direct market access (DMA) possible (given a prearranged agreement with an intermediary to control counterpart risk, an electronically submitted order can be routed through the brokerage firm on its way to market).

The first exchange to demutualize, the Stockholm Stock Exchange, incorporated as a for-profit company in January 1993. Following Stockholm's successful change of its governance structure, many other market centers have followed suit, including NYSE Euronext, the NASDAQ Stock Market, the London Stock Exchange, and Deutsche Börse. In addition, all of the privately owned alternative trading systems in the United States and European markets are for-profit operations.

From a public policy perspective, privatization has delivered its promise of helping to overcome technological inertia. The members in a membership organization are the broker/dealer intermediaries who have traditionally been at the center of the market. As we have noted, this constituency understandably resists any technology innovation that would disintermediate them out of trading. The for-profit organizations are freed from this impediment, and they can more readily raise capital, form alliances, and make acquisitions without resistance from members.

The operations of the demutualized exchanges are centered on their electronic trading platforms. NASDAQ, in changing to a demutualized structure, has distanced itself somewhat from its dealer community and, in so doing, has focused attention on its own electronic platform. Demutualization of the London Stock Exchange went hand in glove with the introduction of its electronic platform, the Stock Exchange Electronic Trading Service (SETS). The NYSE's route to privatization was its purchase of Archipelago, an equity-based electronic marketplace.

But questions can also be raised about the for-profit structure. Doubts have been expressed about the incentive of an equity-based, for-profit organization to effectively perform the self-regulatory obligations that membership organizations have traditionally shouldered. One could also question the relative efficiencies of the for-profit exchange compared to the mutualized market with regard to the provision of public goods-type services involving the quality of price discovery and the provision of liquidity to temper intraday price volatility, a market quality issue that we discussed in Chapter 7.

Ultimately, selection of an exchange's governance structure should depend on the function that the market center is purposed to perform. Before it demutualized, a market center such as the NYSE provided infrastructure. It invested in major electronic equipment for order handling and information dissemination, but it did not operate an execution platform. In the old days, independent broker/dealer and specialist firms (now called designated market makers) on the Big Board's trading floor handled all of the orders and turned them into trades, as they continue to do to a lesser extent in the Exchange's hybrid market system. Currently, the Exchange does operate a trading platform of its own that handles customer (investor) orders. In the past, the Exchange was also heavily involved in self-regulatory activities. This ended in July 2007 when the self-regulatory operations of the NYSE and the NASD (National Association of Securities Dealers) were consolidated in the Financial Industry Regulatory Authority (FINRA), a new non-government regulator for all securities firms.

These two items—SRO operations and the trading floor—gave the NYSE its distinctive brand. Both are natural functions of a membership organization. Without a floor, the NYSE would not need its members; without members, its extensive surveillance operations would take on less importance and/or could be outsourced (as has now been done), and/or could be turned over to a government regulator. Despite its obvious positives, demutualization has weakened the Exchange's branding and perhaps also its dominance in the industry.

## Self-Regulation

Whether it is a for-profit organization or a mutual organization, an exchange runs a *regulated business*.[1] Regulation, organization, technology, systems and logistics are key functions of an exchange. Insofar as it sets, supervises and enforces its own rules, an exchange is a *Self-Regulatory Organization* (SRO). Self-regulation sets the detailed rules for virtually every aspect of the marketplace. An exchange is largely defined by its rule book.

Every market needs basic rules and common practices, along with some regulation, self-imposed and/or externally imposed, to enhance its quality by delivering honesty, fairness, orderliness, and consistency. Securities trading in particular requires considerable regulatory attention, and both the traded instruments (the securities) and the listed companies must be focal points of the regulation. A high degree of standardized disclosure and transparency about listed companies is also a prerequisite for public confidence.

---

[1] Material in this subsection has been drawn from Robert A. Schwartz and Reto Francioni, *Equity Markets in Action: The Fundamentals of Liquidity, Market Structure & Trading* (John Wiley & Sons, 2004).

Stock exchanges are typically among the most regulated micro markets. But regulation also reduces the freedom of market participants. An important maxim is that there should be as much regulation as necessary, yet as little as is possible. Moreover, an optimum combination of government regulation and self-regulation is crucial to ensuring an exchange's competitiveness.

From a broad perspective, an exchange (and its rule book):

- Might be regarded as a "public utility" that has the task of establishing good liquidity, good prices, and low transaction costs.
- Is a place to build market size and efficiency for public companies and its members.
- Creates value by establishing and applying rules and regulations that ensure fair and reasonably transparent price discovery. *Rules* are typically established by the exchange itself (and they commonly need to be formally approved by a regulator); *regulations* are typically established by external regulators.

The overall objectives of exchange rules and regulations are (1) investor protection, (2) system protection, and (3) the achievement of a "fair and orderly" market. To meet these goals, an exchange must have independence and the freedom to operate within its own particular regulatory environment. As with any marketplace, there is always some risk whenever buyers and sellers come together. The goal of an exchange is to reduce this risk without imposing undue administrative burdens on market participants.

The expanding globalization of securities markets is giving rise to *competition in regulation* between national and/or supranational capital markets. What happens if the rules and regulations in one particular marketplace are too burdensome, inefficient, or inconsistent, and therefore are not competitive? Under these conditions, investors (especially institutional investors with the necessary technology, clout, and sophistication) will route their orders to another marketplace where the regulatory structure is more advantageous. This can be done with the use of modern technology that enables an investor, with a simple mouse click, to move his or her orders from one market to another. The process is called *regulatory arbitrage*.

With regulatory arbitrage, an exchange's legal framework is under considerable pressure to be competitive. If it fails, the consequences can be huge. And if an exchange is not successful in the international competition for liquidity, the cost of shareholder equity can rise significantly.

The regulatory functions necessary for an exchange to operate effectively comprise:

**The Primary Market**

- Vetting of prospectuses.
- Admissions to listing and trading.

**The Secondary Market**

- Supervision of trading.
- Market surveillance in real time (for instance, monitoring for price manipulation).
- Market surveillance posttrade (e.g., looking for insider trading by members, intermediaries, and/or customers).
- Establishing general rules of business (including relevant aspects of post-trade clearing and settlement).
- Overseeing member positions and market risk.
- The specification and monitoring of obligations concerning financial reporting and disclosure, and the dissemination of price and other sensitive information.
- Monitoring compliance with corporate governance regulations.

**Membership, the Admission and Vetting Of**

- Member firms and other market participants.
- Traders.

The delegation to an exchange of the competence and power to set its own rules and regs is usually stated in a national or federal law, typically in the form of a "Securities Exchange Act." The law sets the legal framework and general objectives, but delegates the establishment of rules to the exchange. The rules, in turn, must usually be approved by a national or federal securities commission—for instance, the SEC in the United States, the Financial Services Authority (FSA) in the United Kingdom, and the Federal Banking Commission in Switzerland. This form of self-regulation means that rules are not established by the government authorities, but by the private organizations. The principle is widely accepted and practiced in other industries. For instance, in accounting there is the International Accounting Standards Board (IASB), the U.S. Financial Accounting Standards Board (FASB), and so on.

Government regulation and self-regulation should complement each other, not conflict with each other. Effective self-regulation relies on an appropriate interplay between a government's authority to approve and supervise, and an exchange's right as a private organization to define its own rules. Good regulation represents a cooperative effort by the government regulator and the SRO, with the greater power of the former and the superior flexibility of the latter being combined to create the right environment. Finding the appropriate mix of the two is more of an art than a science. It is a customized, tailor-made mix of (1) underlying principles that remain unchanged over a long period of time, and (2) many facets of the market that are in constant development. A rule must be set only when necessary, it must be in line with market needs, and it must be implementable, enforceable, and therefore effective.

The advantages of self-regulation are self-evident. With quickly evolving technology, with product differentiation being a key competitive variable, and with market participants' need for clarity, it is of critical importance that the people who are defining and setting the rules be as close as possible to the key issues. Self-regulation of a marketplace means faster, more flexible and more effective adaptation in a fast changing environment. Being faster than a governmental procedure, self-regulation reduces the potential for crippling *regulatory lag*. This thought deserves emphasis: being able to respond quickly to market developments is of major strategic importance, and the ability to capitalize on and to integrate the know-how of market participants is key.

On the other side of the coin, the disadvantages of self-regulation include the possibility of competition being impaired by cartelized practices, of weaker players in the market being placed at an unfair disadvantage, and of interest groups or stronger players swaying the governmental regulatory process (an act that is referred to as *regulatory capture*). We must conclude that there is no ideal regulatory model, that the proper balance between government regulation and self-regulation must be resolved in the context of the particular market at issue. In achieving the proper balance, it is advisable to keep in mind a dictum that Montesquieu (1689–1755) expressed in his treatise, *De l'Esprit des Lois* (1748): "Useless laws weaken necessary laws."

## TECHNOLOGY, AUTOMATION, AND THE NASDAQ STOCK MARKET

Much technology development involves computers and the automation that computer technology enables. The equity markets are an excellent case in point. There is the power, speed, and capacity with which computers transmit and process information (e.g., prices, quotes, and trade sizes, along with fundamental information about the listed companies). There is the ability to tie together disparate information sets, different marketplaces for the same product (e.g., stocks), and different products (e.g., stocks and options). Computer technology has changed the very structure of the equity markets. From Toronto to Paris to Stockholm, from Deutsche Börse in Germany to the New York Stock Exchange and the NASDAQ Stock Market in the United States, equity markets have been transformed from slow, human-intermediated, floor-based, or telephone-based markets to high-speed electronic trading environments.

## The NASDAQ Stock Market

In the 1970s, computer technology enabled the birth of the NASDAQ Marketplace. More recently, it has enabled NASDAQ's virtually total re-engineering. Let's take a look.

The 1938 Maloney Act Amendments to the Congressional Securities and Exchange Act of 1934 created the National Association of Securities Dealers (the NASD) as a self-regulatory organization (SRO). Headquartered in Washington, D.C., the NASD has primary responsibility for regulating brokers and dealers and, in this capacity, has imposed a uniform set of rules for its members. It also develops and operates the technological infrastructure that facilitates the operations of its members.

The NASD's first major electronic system, NASDAQ, began operations in 1971. NASDAQ at its inception was not an order execution system, but a nationwide electronic network that, for the first time, displayed market maker quotes for NASDAQ issues on terminals in brokerage offices across the country. Prior to NASDAQ, an *over-the-counter* (OTC) market existed which included market makers and other member firms that were linked together by telephone lines. The OTC market itself originated in an era when stocks were bought and sold in banks, and the physical certificates were passed "over-the-counter."

Initially, NASDAQ was not a market, but the NASD's Automated Quotations (AQ) system. Its striking success as a system resulted in its name being applied to the market that it itself had helped to create. The real-time, electronic NASDAQ system had a tremendous impact on the efficiency of the market. It integrated the dealer market, caused spreads to tighten, and improved the quality of price discovery and trade execution.

With success in hand, in 1982 the NASD introduced NASDAQ/NM as its premier market (NM stands for "National Market"). Issues on the NASDAQ/NM list were the largest and most actively traded NASDAQ stocks. Along with the publication of bid and ask quotations, stocks traded in the NASDAQ/NM were subject for the first time to last sale price and volume reporting. NASDAQ/NM market makers were (and still are) required to report transaction prices and sizes within 90 seconds of a transaction's occurrence.

Before the system went live in 1971, dealer firms feared the competition that NASDAQ would introduce, and they rebelled against the idea of displaying their quotes on the NASDAQ screen. But despite having been staunchly resisted at the time, the innovation proved to be a great success for the industry. The configuration of dealer firms in the industry changed, and some firms grew, prospered and, for the first time, gained national prominence. To do business, a dealer firm must receive orders from brokers, and the automated quotation system enabled those firms that make the best markets to receive more of the order flow. With the computerized quote display system, a dealer firm can successfully get order flow and make a market in a stock it has not previously traded simply by registering its intention a day in advance and then posting quotes on the screen.

Prior to the success of NASDAQ, the natural progression for a U.S. company as it grew from a small start-up to a major firm with national stature, was first to trade OTC, then to list on the American Stock Exchange (Amex), and finally to transfer its listing to the NYSE. This changed as NASDAQ first eclipsed the Amex and then challenged the NYSE by retaining many of its premier listings such as Microsoft and Intel.

Until the latter part of the 1990s, companies had a well-defined choice between listing on the NYSE's primarily order driven (agency) market, and on NASDAQ's competitive dealer market, and competition between these two market centers was (and remains) intense. Today, NASDAQ is no longer a predominantly quote-driven environment, the NYSE is no longer a primarily floor-based exchange, and the distinction between the two market centers has blurred. And the Amex no longer exists as a separate exchange; its acquisition by NYSE Euronext was completed on October 1, 2008.

NASDAQ has further reengineered itself to include, along with its traditional dealer market, two nondealer facilities: (1) the NASDAQ Market Center (a seamless, transparent, electronic trading platform for all NASDAQ and NYSE-listed securities that was launched in 2002), and (2) its opening and closing electronic call auctions (these were instituted in 2004 and are referred to as its opening and closing "Crosses"). Today, limit orders and call auctions are playing a prominent role in the NASDAQ market, while market makers, at the same time, continue to be part of the action. According to Robert Greifeld, NASDAQ's CEO,

> *The NASDAQ market model is about competition. We recognize the value of intermediaries, the NASDAQ market makers, in this context. We have over 400 total participants including more than*

> *250 market makers in the NASDAQ stock market. They play a vital role. We need them for the liquidity they provide. We need them for what they bring to price discovery. We need them for the stability that they offer the mid- and small-cap segments of our market and all of our listed stocks during stressful conditions.*
>
> —Robert Greifeld, "NASDAQ Market Makers in the NASDAQ Stock Market," in *The Equity Trader Course*, R. Schwartz, R. Francioni, and B. Weber (John Wiley & Sons, 2006), pp. 263–264

This brief look at the development of the NASDAQ Stock Market shows how a firm and indeed an industry can be transformed by technology innovation. With this in mind, let's take a broader look at the role that automation has played in reshaping the micro markets.

## Automation's Impact: Securities Trading and Beyond

For over a quarter of a century, markets around the globe have been experiencing massive technology change, and much production has been automated.[2] Automation is the substitution of self-operating machinery or electronics for manual or animal effort to support or control a broad spectrum of processes. Examples range from automatic teller machines, to robotic farm tractors, to securities transactions, and beyond.

Henry Ford's use of the moving assembly line to produce Model T Fords in the early 1900s was a precursor to today's assembly lines that feature robotic assembly stations, automated inventory control, and testing and defect detection, all of which can be quickly reconfigured to accommodate variations of car models. Information technology (IT) is a form of automation that is used to process data, transmit information, or handle transactions such as ordering merchandise, buying or selling securities, or making hotel reservations.

Technology change has transformed economic arrangements in a myriad of ways. It has profoundly impacted production processes; it has increased the speed, accuracy, and sheer volume of output; and it has eliminated some kinds of tedious, repetitive work. Automation that extends the reach of information transmission, processing and control generates economies of scale that lead to firms being larger. It also allows production to be carried out in more disparate regions, thereby increasing the intensity of global competition.

---

[2] Material in this subsection has been adapted from Robert Schwartz and Richard Holowczak, "Automation," in *Encyclopedia of Social Problems,* ed. Vincent Parillo (SAGE Reference, 2008), 64–66.

Technology change has far reaching consequences for employment opportunities. Automation generally substitutes for unskilled labor while it complements skilled labor. As such, automation results in the elimination or outsourcing of some jobs, and in the creation of new jobs. As skill requirements change, members of the labor force must be retrained and, across the board, educational demands are raised.

Automation has far reaching impacts on the consumer side. The prices of existing products and services are lowered, their quality is increased, and entirely new products and services such as digital entertainment (CDs, DVDs, MP3s, etc.) are created. Automation has also increased the productivity of consumption within the home, contributing to the quality of leisure time. Examples include dishwashers, microwave ovens, and automatic lawn sprinklers.

Automation generally has positive impacts on productivity, economic growth, and the quality of life. Nevertheless, it is important to stress that technology is neutral. It is a tool that enables tasks to be accomplished. The social desirability of a technology change depends on the tasks to which the tool is put and on just how it is used. Computationally fast computers are great. But what are the market quality implications of lightning fast order handling that gives importance to microseconds? What are the impacts on the quality of stock price determination and volatility of lightning fast trading that has resulted in computer driven algorithms being used along with (and also in place of) "on the spot" human decisions?

Automation typically involves the substitution of prespecified, codified rules for human judgment. This might be fine during normal times, but anomalies can cause breakdowns and, under stressful conditions in particular, human judgment retains its importance. Automation may duplicate human actions with machines and electronic technology, but, when automated, the tasks themselves can change. For example, many secretarial jobs have morphed into administrative functions in an automated office environment. Electronic technology can blur national boundaries and, in the words of Thomas Friedman, create a "flat world."[3] For example, a phone call from New York City to Akron, Ohio, can be routed through India without either party to the conversation knowing it.

Automation can accelerate the speed with which events occur. While it is desirable to get tasks done quickly, speed can have negative consequences when different processes must be coordinated. For example, faster cars are not necessarily safer cars, especially if they are cruising down the highway at great speed. All told, automation gives rise to new public policy issues that challenge government regulators.

---

[3] Thomas L. Friedman, *The World Is Flat: A Brief History of the Twenty-first Century* (Farrar, Straus & Giroux, 2005).

In recent years, technology change has affected many industries. As we got a taste in the previous subsection, one of the most important and complex industries, the securities markets, has been transformed in the past quarter century from a horse and buggy era to a jet age. In the early 1970s, equity trading was, for the most part, a manual, human-intermediated process. Orders were transmitted to brokers who physically met each other (either face-to-face or by phone) to trade shares of a stock. Then, gradually, automation entered the picture. The first role for the computer was to deliver trading instructions and information about the prices at which people were willing to buy and to sell, along with the prices and sizes of realized trades. (Remember, in the United States the NASD introduced its automated quotation system, NASDAQ, in 1971). But information dissemination is not the same as trade execution.

Exchange-based *trading* was itself automated in 1977 when the Toronto Stock Exchange introduced an electronic platform. Over the next twenty years, European market centers, from Sweden to Spain, including London, Paris, Switzerland, and Germany, replaced antiquated floor-based trading with high-tech, electronic systems. In the United States, Instinet (1969) and Archipelago (1997) were among the first to introduce electronic trading. NASDAQ rolled out its own automated trading system in 2002 and, in spring 2006, the New York Stock Exchange (NYSE) initiated its Hybrid Market, an electronic system that coexists with the Big Board's traditional trading floor. As we saw earlier in the chapter, the replacement of manual broker/dealer intermediation with electronic systems that are owned by the exchanges has paved the way for exchanges to convert their organizational structures from not-for-profit memberships to privatized, for-profit entities, and the NYSE did just that (in 2005).

In the securities markets, automation's impact extends further. Electronic technology in the equity markets has enabled the rapid calculation of price indices such as the Dow Jones 30 Industrial Average and the S&P 500. While at first thought this might seem benign, it is in fact of major importance. Virtually continuous information about the continuously changing indices has supported the trading of new financial products including index futures, index options, and exchange traded funds (ETFs). Real-time index values are also valuable pricing guides for individual shares. And there is more: automated trading provides a faster, more error-free transmission of trade data into posttrade clearance and settlement, thereby increasing the efficiency of these two critical operations. The electronic capture of intraday records of all quotes and transactions has also facilitated the overview and regulation of trading, and it has been a boon for industry analysis and academic research.

Nevertheless, when it comes to equity trading, automation has also introduced new challenges. There is an easy part to trading: delivering orders to the market and reporting back quotes and transaction prices. There is a difficult part to trading: handling orders at the most critical time, when buys meet sells and the orders are turned into trades. Trading involves more than simply transferring share ownership from one participant to another at a preestablished price (as is the case when a passenger books an airline seat). Trading also entails finding the prices at which trades are made, a complex process that we have referred to as *price discovery*.

Electronic trading does not totally dominate human-to-human trading. Automating the point of trade for small, retail orders (perhaps 1,000 shares or less) for big capitalization stocks is not difficult. A large trade (perhaps 500,000 shares or more), on the other hand, can be extremely difficult to handle. Time, skill, and risk taking are all required.

The difficulty of handling large orders for all stocks, and all orders for mid- and small-cap stocks, is one reason why automation has proceeded as slowly as it has in the equity markets and why, for instance, it took the NYSE years to offer an electronic platform as an alternative to its labor-intensive trading floor. Moreover, while electronic information transmission is lightning fast and human-to-human information transmission is considerably slower, the human-to-human transmission can include a broader spectrum of thoughts and emotions (a tone of voice or a facial expression can itself convey information). But technological inertia has also retarded the pace of change in market structure. We considered technological inertia as a source of market inefficiency in Chapter 7, and have expanded on the thought in this chapter. This market imperfection has indeed retarded the pace of technology change in the equity markets.

Trading is a very challenging application to automate. For trades to be realized in an equity market, the buy and sell orders of a broad spectrum of participants have to be brought together, and automating this interactive process is far from simple. The software which implements the structure of an electronic market has to be specified in far greater detail than is required for human-intermediated trading and, as the architects say, the devil is in the details. Francioni, Hazarika, Reck, and Schwartz put it this way:

> *For instance, a human agent, or specialist, has historically handled price determination at NYSE openings. This function is performed with reference to various rules, but the specialist is also free to exercise reasonable judgment. Further, human-to-human interactions can evolve naturally as problems, opportunities, and new*

> *competitive pressures arise. In contrast, with a fully electronic opening, every possible condition that can occur must be recognized and a rule for dealing with it specified; and electronic interaction can be changed only by rewriting the code that specifies with step-by-step precision just how orders are handled and turned into trades and transaction prices.*
>
> —Reto Francioni, Sonali Hazarika, Martin Reck, and Robert A. Schwartz, Francioni, R., Hazarika, S., Reck, M. and Schwartz, R., "Equity Market Micro Structure: Taking Stock of What We Know," *Journal of Portfolio Management,* Fall 2008, pp. 57–71.

So, here is the big picture. Electronic environments generally differ a great deal from human-to-human environments (either telephone connected or face-to-face), and replacing people with machines and computers is not a simple matter of having machines and computers take over tasks that were previously performed by humans (and building modern roads is not, as the expression goes, equivalent to paving the cow paths). In markets around the world, human agents have staunchly resisted the introduction of electronic technology that, by automating activities, eliminates otherwise profitable jobs.

In equity trading, automation has driven down commission costs and, in response, volumes have exploded. Orders are being delivered and trades executed within tiny fractions of a second. However, the sequence in which orders arrive remains important, and subsecond time frames are of no substantive importance per se. Concurrently, large block orders are commonly being shot into the market in protracted sequences of smaller tranches over protracted periods of time (perhaps two hours or more). This practice of "slicing and dicing," and the speed with which events can happen in the automated environment, have pressured traders to use new computer tools to time, size, price, and otherwise handle their orders. The automated, rules-based procedures, referred to as *algorithmic trading*, are both a product of automation and a symptom of the complexities that electronic, high-speed trading can introduce.

In early 2009, considerable interest emerged in the financial press about high-frequency trading that powerful electronic technology has enabled. A *New York Times* article by Landon Thomas Jr. on September 4, 2009 (pages B1 and B6) described how Optiver, a trading company based in Amsterdam, using its own capital, " . . . seeks to profit on razor-thin price differences—which can be as small as half a penny—by buying and selling stocks, bonds, futures, options, and derivatives." The article goes on to say that Optiver's business model ". . . deploys a sophisticated software system called F1 that can process information and make a trade in 0.5 milliseconds—using complex algorithms that let its computers think like a trader." This state-of-the-art technology, in the hands of the cutting-edge firm Optivar has resulted in fractions of a cent and fractions of a fraction of a second turning into very handsome profits. But what does this say about random walk, market efficiency, and fairness? Questions have been raised in the press, on the part of regulators, and by others. New technology can clearly give rise to new problems.

Nevertheless, as we all know, technology offers tremendous promise and, driven by automation developments, economic activities around the world have been impacted. Automation that enables people from disparate parts of the globe to access a market and communicate virtually instantly with virtually equal facility has flattened the world of trading and commerce. Technology is indeed a powerful tool.

It is also a harsh taskmaster. Throughout history, and even today as seen in the equity markets, automation's introduction has rarely escaped controversy. Its transformative power disrupts the status quo and creates new sources of friction even as significant reductions in time, effort, and mistakes are realized. As technology transforms economic processes and alters the distribution of wealth and power across economic agents, regulatory intervention must be reshaped to keep up with the evolving dynamics of the micro markets. This is not a simple task. We turn to governmental regulatory intervention in the next section.

## THE EVOLVING REGULATORY FOCUS

Regulation has had, and will continue to have, a profound impact on the equity markets. Much debate has attended each major regulatory initiative regarding market structure, starting with the Securities Acts Amendments of 1975. We can all be sympathetic to the regulatory goal of making our markets fairer and more efficient. However, given the enormous complexity of the economic issues involved, many students of the market have taken a more free-market stance with regard to the evolution of market structure. But there is one way in which the interaction between technology and regulation can be very beneficial. Life may simply be too cushy for a market center that faces only weak competition, and a market center with monopoly powers may be too slow to adopt a new technology that would increase its efficiency. Consequently, regulation can be employed to shake things up.

The very *threat* of regulatory intervention can and has been used to good advantage. Here is a case in point. In the late 1970s, the New York

Stock Exchange had an order-consolidation rule on its books, NYSE Rule 390. We will consider this rule in more detail shortly but, in brief, it required that NYSE member firms bring customer orders for exchange-listed stocks to an exchange to be executed. While the requirement was viewed by many as an unjustified monopolistic constraint on competition, others in the industry adamantly defended it. The industry downright feared its removal, and this fear enabled the SEC to use the *threat* of removing Rule 390 as a club that got the industry to move forward with the development of a better-integrated National Market System. We return to Rule 390 later in this chapter.

Regulatory intervention along with competition can trigger the reengineering of a market. Reengineering has occurred throughout the equity markets of Europe, starting in 1986 with London's Big Bang, a government-instigated event that totally altered the British market (among other things, it abolished fixed commissions). In the United States, NASDAQ and the NYSE have both reengineered themselves. A major U.S. regulatory initiative in 1997 (the introduction of new order-handling rules) ushered in a period of intense competition, followed by a period of reconsolidation. Why the reconsolidation? Because the equity markets exhibit huge economies to scale, and enormous network externalities. Presumably, after reconsolidation, some old problems have been solved. But will the new regulations have "unintended consequences"? Will new problems appear? Will these lead to fresh calls for further market structure regulation that will be followed by a further round of fragmentation, intense competition and then, once again, reconsolidation? Will our markets in the future be more efficient than they are today? Will we see yet lower transaction costs and more accurate price determination? Will a better, more effective and more efficient regulatory structure have been put in place? Not one of these questions can be answered with a simple yes or no.

In Chapter 7 we focused on the innate efficiencies and inefficiencies of micro markets. When operating in an environment that is free, frictionless, and competitive, a micro market can indeed deliver very desirable results from a public policy point of view. But real-world markets are not entirely free, they are far from frictionless, and they are not always competitive. Despite imperfections in the political process, regulation is needed and, in the securities markets, it has grown substantially, starting with the Securities Acts of 1933 and 1934. As we noted earlier in this chapter, the regulatory goals for the securities markets are threefold: (1) assure an honest market, (2) enhance efficiency with regard to the provision and pricing of broker/dealer services, and (3) enhance efficiency with regard to the pricing of shares traded. Let's take a closer look at these goals.

### Honesty

We offer the following dictum: To operate effectively, a market must, first and foremost, be honest. Achieving an honest market includes guarding against the manipulation of information, prices, and trading volumes, and preventing other abuses of power and position by insiders and other professionals.

With this objective in mind, major federal regulation of the securities markets began with the Securities Act of 1933 (sometimes called the "Truth in Securities Act") and the Securities Exchange Act of 1934. As we will discuss shortly in further detail, the primary motivation for this legislation was to ensure reasonable information disclosure and to prevent dishonesty and market manipulation. The acts were designed to protect the ignorant from being exploited by those who might distort reality and entice the gullible into undesirable transactions.

The U.S. Securities and Exchange Commission (which was established by the 1934 Act) has been particularly concerned with the fairness of the information flow, and the agency has devoted major resources to preventing abuses of power by insiders. For those involved with the activities of a corporation (with positions ranging from top management to the proofreader in a print shop), information is a freely generated by-product of some other activity. A primary objective of insider trading regulations is to prevent those for whom information is a free by-product from making excess trading profits at the expense of others who do not have a fortuitous position vis-à-vis the information flow.

On August 15, 2000, in a further attempt to promote full and fair disclosure of new information, the U.S. Securities and Exchange Commission adopted Regulation FD. Reg FD requires that any information that has been provided to some stock analysts and/or to any other securities market professionals who may themselves seek to trade on the basis of that information, must be disclosed, promptly, to the entire marketplace, rather than to only a selected group of individuals.

Laws are one thing; enforcing them is another matter. Given the complexity of the securities markets and the vast sums of money involved in trading and investing, policing the markets is an enormous task. Since the markets' inception, manipulators, rogue traders, and other criminal types have sought to profit illegally. One trick, a *Ponzi scheme*, is named after Charles Ponzi, who in the early years of the twentieth century was the first known person to use it in the United States. A Ponzi scheme is a pyramid scheme that uses funds raised from new investors (rather than portfolio profits) to credit existing investors with handsome returns. As the investor base expands, however, increasing amounts of new money have to be raised to perpetuate high returns for existing investors. But the investor base cannot expand forever. After a while, the scheme collapses under its own weight, and duped participants lose their investments.

On December 11, 2008, a highly prominent member of the U.S. financial community was arrested for having perpetrated the largest Ponzi scheme on record. Bernard L. Madoff had been a pioneer in electronic trading and market making. He had a powerful influence with the SEC regarding market structure regulation, and he had served as Chairman of the Board of the NASDAQ Stock Market from 1990 to 1993. Along with his firm, he also managed a portfolio. With the stellar reputation of the firm that carried his name, with his persona, and with his professional standing in the industry and broader community, he was able to attract an ever-expanding array of investors that included wealthy individuals, hedge funds, charitable organizations, universities, hospitals, the International Olympics Committee, and so on. His base grew to include investors from the United States, Europe, the Middle East, China, Mexico, South America, and so on. At the time of his arrest, he himself put the loss at \$50 billion.[4] Instantly, his name became known and vilified throughout the United States and internationally. Many people who knew him were incredulous, and the same question was asked repeatedly: How could this have happened?

The cataclysmic event is a stark reminder of how important honesty is to the effective working of a free market, and of how difficult controlling dishonest behavior can be. With regard to this one, we note that it was a roughly 40 percent drop in the U.S. equity market that occurred in 2008, not the regulatory watchdogs, that brought this terrible infraction to light.

## Efficiency of the Market for Broker/Dealer Services

The second regulatory objective is to enhance market efficiency with regard to the provision and pricing of broker/dealer services. This aspect of efficiency entails keeping commissions and spreads competitive, and ensuring the financial responsibility of broker/dealer firms.

Concern about the competitive efficiency of the markets developed with the back-office crisis of the 1960s. The period 1963–1968 was called the go-go years. After a collapse of the market in 1962, the economy had entered a strong expansionary phase and, with it, trading soared. The industry was not prepared for this. Back-office staffs in the 1960s were small, and procedures were primitive. The number of trained people was insufficient, and new systems often failed because of installation and programming difficulties. Stock certificates were misplaced, and some were even found in the back of filing cabinets, tucked behind water pipes, in trash baskets, and in various other inappropriate places. A substantial increase in brokerage firm failures ensued, as the houses could not keep up with the paperwork, and as the number of *fails to deliver* (either certificates for shares that were sold or payments for shares that were bought) mounted.

---

[4] As of April 2009, the amount lost was estimated to be \$65 billion.

In an attempt to deal with the crisis, the markets were closed on Wednesdays and daily trading hours were shortened to give back-office staffs time to catch up. These efforts, however, proved insufficient, and the way was paved for new market regulations. The primary regulatory responses were to enforce minimum net-capital requirements and to increase the size of special funds for investor protection. These changes were first made by the NYSE in its SRO capacity, but the SEC perceived the supervision to be inadequate.

As the agency grew more concerned, Congress was drawn into the situation. In 1970, Congress passed the Securities Investor Protection Act and, in so doing, established the governmental insurance of customer accounts (SIPC, the Securities Investor Protection Corporation). Until that time, Congress had been focused on preventing fraud, manipulation, and other abuses of power. Now the picture changed. Government, having undertaken the role of insurer, for the first time had a vested interest in controlling the industry. For self-serving reasons alone, they wanted to prevent the continuing failure of broker/dealer firms. And so Congress came to be involved in the competitive structure of the securities markets.

Another impetus was afoot. In the early 1970s, commissions charged for handling large institutional orders were exorbitant, and the marketplace was increasingly fragmenting as institutional traders were sending their orders to competing markets (such as the regional exchanges) in their efforts to avoid these rates. Related public policy issues concerning the competitive efficiency of the markets included the absence of spatial integration among various parts of the industry, the quasimonopoly position of various market makers (primarily the stock exchange specialists), and restrictions on entry into certain market centers (primarily the New York Stock Exchange and American Stock Exchange).

Congressional concern about the competitive efficiency of the market for broker/dealer services resulted in the enactment of the Securities Acts Amendments of 1975. That landmark legislation had two provisions of particular importance: (1) It precluded the securities exchanges from imposing fixed commissions, and (2) it mandated the development of a National Market System (NMS). We return to both of these later in the chapter.

### Efficiency with Regard to the Pricing of Shares Traded

The third regulatory objective is to enhance market efficiency with regard to the pricing of shares. This aspect of efficiency entails improving the market's architecture so as to support adequate liquidity provision, to contain accentuated volatility and, in general, to facilitate price discovery.

The SEC's sister agency, the Commodities Futures Trading Commission (CFTC), has long recognized price discovery as a major function of a securities market, the reason being that prices discovered in the futures markets are commonly used for price determination in the associated cash markets. Indeed, a major economic rationale for the futures markets is their price-basing function. For instance, the price of cocoa is believed to be more accurately determined by the meeting of many buyers and sellers in the market for cocoa futures than by the meeting of a far smaller number of firms that buy and sell the cocoa itself. Therefore, cocoa traders look to the futures markets to set their spot prices.

The quality of price discovery, which we considered in some detail in Chapter 7, is an important public policy issue. Clearly, a 25¢ discrepancy in the price of a stock far exceeds the brokerage commission on a transaction. And there are further costs of inaccurate price discovery. Prices established in the major market centers signal information to traders and are used for marking-to-market, for the maintenance of margin accounts, for price basing in related markets, and for the evaluation of estate held shares, equity mutual fund shares, and so on.

Efficient price discovery is the most difficult of the three regulatory goals to comprehend and to implement. It is an objective that very much pertains to the attributes of the equity markets but, as we suggested in Chapter 7, it may also have applicability for a considerably wider array of micro markets. Given the importance of accurate price discovery, more attention should be paid to this objective.

## GOVERNMENTAL REGULATORY INTERVENTION

Political and market processes differ in several important ways. The political process involves a different incentive structure than the private market: first and foremost, the maximization of electoral votes rather than the maximization of financial wealth. In the legislative process, votes can be used as a medium of exchange, whereas money is the medium of exchange in the market economy. And political power is derived from the ability to win electoral votes, while economic power is the command over resources, both real and financial. Moreover, the distribution of power across participants in the political process differs from the distribution of power in the market process. Some who are powerful in the private market are not in the political arena, and vice versa. The political process typically admits outsiders whose main considerations include not only the best interest of a particular industry, or the pros and cons of a particular issue, but also a broader set of pressures to serve other constituencies as well.[5]

In our discussion in Chapter 7 of the efficiency of competition, we set forth a number of reasons why the free market may fail to deliver fully efficient results. In parallel, we now consider various reasons why a government's regulatory intervention may itself be flawed.

**Inadequate procedural safeguards:** Procedural safeguards in the governmental process differ from those implicit in the market process. The private market itself is a strict and harsh regulator—those who do not operate effectively and profitably typically do not survive (unless bailed out by government). Unfortunately, government officials who do not regulate well may nonetheless survive as regulators. Instituting effective procedural safeguards is at best costly; at worst, it may be impossible.

**Jurisdictional disputes:** The governmental process can involve competition between rival government authorities for regulatory turf. For instance, the growth of futures and options trading in the early 1980s caused confusion as to which agency, the SEC or the CFTC, would have jurisdiction over these new products. Likewise, change in both the securities and banking industries has generated territorial disputes between the SEC and the bank regulators.

Regulated firms typically have some freedom to choose their regulators, and regulators typically compete for their constituency. This competition can affect the decisions regulators make, may prejudice the outcome of the political process, and diminishes one's confidence that government intervention will enhance the economic efficiency of a regulated industry.

**The cost of regulating:** Regulation is costly, and government funds are limited. The fact that government funds are allocated by a budgetary process embedded in a political environment further suggests that funds may not be distributed efficiently across competing government projects.

---

[5] For further discussion, see George Stigler, "The Theory of Economic Regulation," *Bell Journal of Economics and Management Science,* Spring 1990.

**The cost of being regulated:** Regulation can impose a heavy compliance cost on a regulated market. Following the Enron scandal that was revealed in 2001, The Sarbanes-Oxley Act of 2002 (which is also known as the *Public Company Accounting Reform and Investor Protection Act* of 2002 and is commonly called "*SOX*") was designed to improve the quality of financial reporting, independent audits, and accounting services for public companies. The sheer amount of paperwork required by the act has proven to be very burdensome and, because of this, some U.S. firms have taken their IPOs to non-U.S. markets (primarily London).

The *inconstancy* of government rules, regulations, and tax schedules imposes further costs, as it generates uncertainty and requires that tax, business, and personal planning be revised.

There is one more critically important consideration to recognize. The extended time that it commonly takes to receive government approval for a new facility or rule change can be extremely costly to a firm that is operating in a fast-moving, highly competitive environment.

**Imperfect knowledge:** More than some may acknowledge, regulators may not adequately understand the problems encountered by an industry, and they may not know best how to develop and institute reasonable solutions. In recent years, both the SEC and the CFTC have solicited opinions, held hearings, and deliberated long and hard about certain critical decisions. However, much remains that is not known about the nature of competition in the equity markets. The problems concerning price discovery, for example, are not easily perceived and, when perceived, are thorny to resolve. All told, the boundaries of knowledge are among the toughest limitations faced by those who seek to govern the market process.

Having taken an aerial view of problems concerning government intervention in the workings of the micro markets, we continue to focus on the securities industry to understand more concretely how the U.S. government has intervened in the development of this micro market, to gain further insight into the complexities of the issues involved, and to draw some conclusions about the effectiveness of the regulatory initiatives.

## REGULATION OF THE U.S. EQUITY MARKETS

The U.S. economy experienced booming prosperity in the 1920s and, as the stock market soared to new heights, a giddy feeling swept over much of the land. But the environment was wild and unregulated, and the widespread prosperity that it brought proved to be unsustainable. In the fall of 1929, the roof caved in. After closing at 305.85 on October 23, the Dow Jones Industrial Average took a wild ride on Thursday, October 24 (it experienced an 18.0 percent swing on that day before closing down 6.13 percent). Monday, October 28, brought a precipitous decline (12.8 percent), and then there was October 29, the day that came to be known as "Black Tuesday," a day that historians often use to mark the start of the Great Depression (the Dow dropped another 11.7 percent on October 29 and closed at 230.07). That was not the end: The decline continued. As of the close on Wednesday, November 13, the Dow had sunk to 166.15, a dismal 45.67 percent lower than the 305.85 it had been at on October 23.

Responding to the Great Depression, the U.S. Congress passed two acts that have established the broad structure for regulating the U.S. financial markets: the Securities Acts of 1933 and of 1934. In this section of the chapter we briefly survey the scene in the 1930s, and then fast-forward to discuss seven major U.S. market structure initiatives that have subsequently been taken, starting with the enactment by Congress of the Securities Acts Amendments of 1975.

The October 1929 crash, the Great Depression, and the Congressional acts that followed underscore the reality that a "free market" requires an effective regulatory structure within which to operate. Here's an analogy: as drivers we value the freedom to pick our routes and the times when we will travel them but, of course, traffic lights, speed limits, and stop signs are a necessity. It would be hard to believe that any free-market thinker would ever suggest unplugging traffic lights and telling the courts not to enforce the payment of traffic tickets. A system of lights and regulations provides a broad-based framework for us all. Nevertheless, only a fine line may separate a regulatory framework from an overly detailed government involvement in the decisions of firms and household. Where should that line be drawn? There clearly is no definitive answer to this one, but we hope that our discussion in this section of the chapter will help to sharpen your thinking about the matter.

### The Scene in the 1930s

Leading up to the crash on October 29, 1929, the markets were indeed wild and woolly. They were almost totally unregulated, manipulation and other abuses of power and position were rife, and very little sunlight filtered down to illuminate the deals that were being made, the transaction prices that were being set, or the corporate books that contained much of the information that, if broadly known, could translate into more appropriate corporate valuations. In a very specific way, monopoly was a major problem. For this one, think not of monopoly in the production of goods and services, but of monopoly in the possession and use of information concerning the listed companies and the trading of their issues. At that earlier time, a few people could be in possession of the inside scoop while many were left uninformed, and some of the uninformed were being duped by the spreading of false rumors. Woefully lacking were reasonable transparency and the enforcement of appropriate rules and regulations that were key to earning investor confidence. The 1933 and 1934 acts sought to rectify the problem, and the results were positive.

Two major building blocks of the regulatory structure were put in place by the 1933 Act (also known as the "Truth in Securities Act," the "Banking Act," and "Glass-Steagall"). First, to control the inappropriate use of inside information, commercial banking was separated from investment banking (this was the Glass-Steagall component of the act). Second, the act created the Federal Deposit Insurance Corporation (the FDIC) to insure bank deposits. The 1934 Securities and Exchange Act put one other major building block in place: It established the Securities and Exchange Commission (the SEC) and charged it with the responsibility of regulating the exchanges and enforcing the new laws. Later in the decade, an additional piece of legislation introduced a fourth building block: the 1938 Maloney Act Amendments of the 1934 act. As we have previously noted, the Maloney Act established the NASD as a self-regulatory organization, and brought the over-the-counter market under control.

The legislative structure that was put in place in the 1930s was badly needed, and the SEC earned a stellar reputation for its effective implementation and enforcement of the new laws. For the first time, listed companies had to disclose their financial statements, the financial statements for the first time had to be independently audited, and the new regulations were effectively enforced. As the historian Thomas K. McCraw wrote,

> *Over the next four decades, the SEC built a reputation as the most effective of all federal regulatory agencies. It was respected and feared by nearly everyone involved in the trading of stocks and bonds, the issuance of new securities, and the governance of corporations.*
>
> —Thomas K. McCraw, "Regulate, Baby, Regulate," *New Republic* (March 18, 2009), pp. 16 and 17

And so, much needed transparency, fairness, and orderliness were brought to the markets in the 1930s, and the effective monopoly on much information was broken.

### The End of Fixed Commissions

The enactment of the Congressional Securities Acts Amendments of 1975 marked the first major government mandated change in the more detailed operations and structural organization of the securities industry in the United States. The 1975 Amendments (1) precluded the fixed commission structure as of May 1, 1975, and (2) mandated the development of a National Market System (NMS). In this subsection we focus on the first. At the time that the legislation was passed, there was much concern about the effect the change to negotiated commissions might have, and some people thought it would be disastrous. Today, the deregulation is widely accepted as having been a success.

Commission income accounted for roughly half of the gross income of member firms of the NYSE in the period 1971 to 1974. As one would expect, broker/dealer firms competed fiercely for the revenue. Brokerage firms offer a package of services—order handling, record keeping, custodial services, advisory services, dividend collection, and, last but not least, research. In the era of fixed commissions, commission dollars alone typically paid for the entire package. The components other than order handling were in essence a rebate to customers in the form of services, rather than in hard dollars.

Therefore, a brokerage firm was able to lower the price it charged for order handling by bundling services (a procedure that is also referred to as *commission bundling*). A major ancillary service included in the brokerage package was, and still is, research. The array of research reports may not have been as valuable as direct dollar rebates, but as part of the package the reports were a free good and, as a free good, they were accepted.

The Securities Exchange Act of 1934 had exempted the NYSE from certain statutes in the U.S. antitrust legislation and, until the 1975 Amendments, the Exchange had been free, subject to permission from the SEC, to set minimum commission rates on stock transactions. In 1968, the NYSE appealed to the Commission for a rate increase. To the Exchange's surprise, the U.S. Department of Justice intervened by presenting to the SEC a brief that not only questioned the need for the requested rate increase, but that challenged the very existence of fixed minimum rates. This brief set into motion the events that led to the total preclusion of fixed commissions on May 1, 1975.

Three undesirable consequences of the fixed commission structure had become increasingly apparent with the growth of institutional trading:

1. The level of the minimum rates was excessively high (purportedly, the excess portion of commissions in relation to the cost of order handling ranged as high as 90 percent for large orders).

2. The market was being fragmented by large traders turning to regional exchanges and to the third market to escape the fixed commission structure on the national exchanges. The third market refers to trades in exchange listed issues that take place off the exchange floor with the aid of a broker.
3. Ancillary services (research, and so forth) were being oversupplied.

After the SEC had opened its investigation of the issue, the burden of proof lay with those who sought to retain the fixed commission structure. The NYSE, as leader of the defense, advanced four main arguments:[6]

1. **Destructive competition in the brokerage industry:** The Exchange argued that because of economies of scale in brokerage, more efficient brokerage houses would drive less efficient houses out of business in a fully competitive environment. The elimination of fixed (high) rates would therefore lead to increased concentration in the brokerage industry. According to the Exchange, the price of brokerage services would then actually be higher for customers because of the enhanced market power of the large firms.
2. **Price discrimination:** The Exchange argued that competitive rate setting would cause the price of brokerage services to be lower for large traders than for small traders because of the greater bargaining power of the institutional traders. Therefore, prices would not be proportionate to costs for traders of different size, and the price discrimination would unfairly disadvantage small investors.
3. **Ancillary brokerage services:** The Exchange argued that fixed minimum commissions were required to ensure that certain essential services other than order handling be provided by the brokerage houses. The primary concern expressed by the Exchange was that the quantity and quality of research would be impaired if competition were allowed to drive commissions to a level that just covered the cost of order handling.
4. **Market fragmentation:** The Exchange argued that fixed minimum commissions were needed to keep the market from fragmenting. Under the fixed commission structure, member firms were given a rate discount for orders brought to the exchange floor. The NYSE believed that this pricing advantage provided the incentive needed to retain its membership; that without fixed commissions it would lose membership and the order flow would fragment; and that with fragmentation various exchange services such as market surveillance could not be adequately provided.

---

[6] The arguments presented by the Exchange, along with an assessment of their validity, have been set forth in West and Tinic (1971) and Tinic and West (1980). This subsection draws heavily on both of these papers. See footnotes 7 and 9.

What do you think of these arguments? How valid are they in light of economic theory, the realities of the marketplace, and history? With regard to the first argument, might commission rates ultimately be higher if big firms were allowed to drive weaker firms out of business and then impose noncompetitive prices? This could indeed occur if average costs for a securities firm are negatively related to volume because of economies of scale (as we have discussed in Chapter 6). It was not inappropriate, therefore, for the NYSE to attempt to establish empirically that average costs do fall with firm size in the brokerage industry.

However, what would the advisable economic policy be if economies of scale existed in brokerage? The appropriate regulatory solution in such a case would be to allow only a few firms (the biggest and the best) to be the providers of the service and then to regulate those firms. That is, rather than establishing a minimum commission at a high price to ensure the existence of inefficient firms, a maximum commission should be stipulated at a low value to keep the price of brokerage in line with the marginal cost of providing the service (take another look at Exhibit 6.5). The economies of scale argument, therefore, did not support the Exchange's case.

Furthermore, the Exchange's empirical findings with regard to economies of scale in the brokerage industry have been challenged by a number of subsequent studies[7] and, in fact, an alarming increase of concentration did not occur in the brokerage industry in the years after the deregulation in 1975. Firms of varying sizes have coexisted, and specialty firms have found their niches in the industry. According to Tinic and West (1980), there is no evidence that the elimination of fixed rates has enabled larger brokerage firms to improve their relative positions in the industry. Let us turn, therefore, to the Exchange's second line of defense: the price discrimination argument.

Because the fixed commission structure itself clearly implied price discrimination (against the large traders), price discrimination was a tenuous argument for the Exchange to advance. Furthermore, with negotiated rates the price discrimination argument does not hold on theoretical grounds if sufficient competition exists between the brokerage firms. The reason is that, with competitive pricing, prices are set in relation to costs, irrespective of demand elasticities in different market segments (as we discussed in Chapter 6).

During the period from April 1975 through December 1978, commission rates did fall appreciably for institutional orders of all sizes and for the larger orders of individual traders. Rates for individual orders in the 200 to 999 share range remained essentially stable, however, and rates actually increased for individual orders in the 1 to 199 share range. Was this realignment of the rate structure the consequence of differential demand elasticities and price discrimination against small traders? Given the evidence that the brokerage industry was reasonably competitive and that there were economies of scale in order handling,[8] these adjustments were more likely to have been attributable to cost relationships and to the elimination of price discrimination against large traders.

---

[7] See R. West and S. Tinic, "Minimum Commission Rates on New York Stock Exchange Transactions," *Bell Journal of Economics and Management Science* (Autumn 1971), for analysis and further references.

The NYSE's third line of defense focused on the provision of ancillary services. Whether bundled or not, the ancillary services would in theory be provided in optimal amounts in a competitive environment so long as no cause of market failure was operative. A market failure argument is possible, however, because the production and use of one of the services, research, may involve externalities. The reason is that market prices depend on the information set. But the prices, in turn, signal information to traders and to firms whose securities are traded, and they are used for price-basing in related markets. Therefore, a decline in the quantity and quality of information production may distort security prices and so be unacceptable to the community in aggregate.

Would a sufficient amount of information be produced as a private good in a free market? Was information being overproduced when commission rates were fixed? These are difficult questions to answer because the empirical evidence is fragmentary and conjectural. All told, however, there is little indication that either the quantity or the quality of research has diminished appreciably with the introduction of negotiated rates. Furthermore, individual investors now have the freedom not to obtain and not to pay for research that is not desired. As Tinic and West indicate,[9] "Under fixed rates and the bundling of services, many investors 'consumed' research for which they would not have paid hard cash. If they were now unwilling to buy these services outright, who would want to say that this was bad?" In any event, following the elimination of fixed commissions, something happened that has surprised many observers of the market—the practice of commission bundling has not ceased.

---

[8] Economies of scale in order-handling cause costs per share to be lower for bigger orders, but do not necessarily imply that average costs for a brokerage firm decrease as its overall trading volume expands, *ceteris paribus*.

[9] S. Tinic and R. West, "The Securities Industry under Negotiated Brokerage Commissions: Changes in the Structure and Performance of New York Stock Exchange Member Firms," *Bell Journal of Economics and Management Science* (Spring 1980), pp. 39–40.

The Exchange's fourth argument, that the elimination of fixed commissions would cause the market to fragment as the Exchange lost members, was totally unsupportable. The market fragmentation that gave impetus to the deregulation was in fact attributable to the fixed commissions. That is, rather than holding the Exchange's membership together, the umbrella of fixed commissions created unjustifiably high rates that drove institutional orders away from the national exchange markets.

The Exchange never did need fixed commissions to hold its membership together. An important reality of the equity markets is that they exhibit strong network externalities, that order flow attracts order flow. *Ceteris paribus*, bigger market centers operate more effectively because they are bigger. The NYSE had failed to appreciate this. The Exchange lacked confidence that its attractiveness as a market center was due to the orderliness of the markets that it provided. The reality is, the NYSE never needed a fixed commission structure to survive. The Big Board lost neither members nor order flow in the years following the introduction of negotiated commissions. Moreover, the demand to trade has proven to be highly elastic with respect to the cost of trading: As commission rates have fallen, trading volume has grown exponentially.

We now have had more than three decades of experience with negotiated commissions. The deregulation, for the most part, did cause rates to fall appreciably. The changing rate structure has benefited the larger investors and, with the advent of discount brokerage houses, has benefited many retail customers as well. There is no evidence that the quantity or quality of research has declined, and individual investors are now free to select the package of brokerage services that they prefer. The change from fixed to negotiated rates, on many counts, must be deemed successful. An equivalent (and also successful) elimination of fixed commissions was successfully instituted in England in October 1986 (as we have noted, the British deregulation is referred to as the "Big Bang").

In the early months of 1975, the advent of negotiated rates had caused great concern in the industry. But reasons were not then, and have not since, been advanced that would justify this concern on the basis of the market failure arguments that we considered in Chapter 7. Nevertheless, competition between alternative trading systems have driven commissions to appreciably reduced levels and, to a large extent, trading services have been commoditized.

The bad news is that a commoditized environment is not very profitable for the suppliers of trading services. In the face of shrinking profits, one might expect a continuing move toward consolidation and who knows, commissions could start to rise again. Interestingly, this chain of events would have overtones of the NYSE's first argument; namely, that unfixing commissions would lead to destructive competition in the brokerage industry. Additionally, lower commission income induces intermediaries to intensify their efforts to profit from proprietary trading, a development that puts them in a more adversarial role (rather than facilitator role) vis-à-vis their customers. Thus the regulation, while generally accepted as desirable, is not without its negative, unintended consequences.

## Best Execution

Best execution has been a holy grail in the United States since the enactment of the Securities Acts Amendments in 1975. In mandating the development of a National Market System, the 1975 Amendments stated, as a goal, that investors' orders be provided the opportunity to be executed, consistent with efficiency and *best execution*, without the participation of a dealer.

At the time the Amendments were passed, institutional participation was far less than it is today, and the best execution requirement was fashioned primarily with regard to retail order flow. More recently, however, attention has turned to institutional investors, a group for whom the requirement is appreciably more difficult to fulfill or even understand. As a consequence of the greater difficulty of specifying best execution criteria for large orders, the concept of best execution for institutional investors is now being applied more to the quality of investment/trading *procedures* than to the costs, per se, of specific transactions.

Best execution has gained attention on both sides of the Atlantic. In London, publication of the *Myners Report* in March 2001 led to a protracted debate about (1) whether traditional fund management contracts give managers adequate incentives to minimize transaction costs (in light of our discussion in Chapter 7, you might recognize this as a principal-agent problem) and (2), if not, what to do about it.[10]

Let's focus on the more fundamental question: What does the term "best execution" mean? The U.S. Congressional Act did not say, and a widely accepted empirical definition has not since been developed. The problem is multifaceted. First, market impact costs and opportunity costs are difficult to measure with any kind of precision on a trade-by-trade basis.[11] Second, good benchmarks for assessing trading performance are difficult to define. Additionally, different kinds of orders require differential handling, depending on the needs of a trader, the size of an order, and the liquidity of the market for the shares being traded. In other words, the execution that is "best" depends on the particulars of the specific case at hand. Further, how does one measure best execution for orders that are broken into smaller pieces for execution over an extended period of time? And how does one specify a common set of best execution procedures to apply to a broad, diverse population of participants?

---

[10] See P. Myner, "Review of Institutional Investment: Final Report," *HM Treasury, London*, March 6, 2001. For further discussion, see R. Brealey and A. Neuberger, "Treatment of Investment Management Fees and Commission Payments: An Examination of the Recommendations Contained in the Myner's Report," *Fund Managers Association*, October 2001, and Neuberger's discussion in Schwartz, Byrne, and Colaninno, eds., *Call Auction Trading: New Answers to Old Questions* (Kluwer Academic Publishers, 2003).

The ambiguities involved in assessing trades with regard to best execution criteria have not discouraged regulatory authorities from pursuing the objective in an attempt to ensure quality executions for public participants. The Financial Services Authority (FSA) in the United Kingdom released a Discussion Paper that reviews policy and invites public comments on the matter.[12] The document presents a broad array of questions concerning the implementation of a best execution requirement.

Our own discussion of best execution is focused largely on the impracticality of applying the criterion to institutional trades. Interestingly, the FSA (much as the Securities and Exchange Commission in the U.S.) is primarily focused on best execution as a "consumer protection tool," where "the consumer" is a retail customer. For retail customers, the assessment of best execution, for the most part, fits into a "snapshot" approach: namely, best execution is judged by matching a transaction price against other prices that exist at the time a trade is made.[13] Overlooked in the snapshot is the timing of order placement, including the submission of large institutional orders in smaller tranches over extended periods of time.

Although a focus on the retail customer is understandable, best execution criteria are applied to all trades. Indeed, it would present another level of complexity to stipulate just what trades are, and what trades are not, subject to a best execution requirement. Consequently, if best execution criteria are not implementable for institutional orders, one might question the advisability of imposing the requirement for any orders. A reasonable alternative may simply be for firms to disclose to their customers the procedures they follow when placing their orders (this was suggested in the FSA's discussion paper), and to let competition take care of the rest. This thought is also echoed in the United States by a report on best execution of the Association for Investment Management and Research (AIMR).

---

[11] More meaningful measures can be obtained by averaging measurements over a substantial number of trades.

[12] Financial Services Authority (FSA), "Best Execution," Discussion Paper (London, April 2001).

[13] In reviewing current policy on best execution, the FSA's Discussion Paper (FSA [2001]) states, ". . . when dealing in securities traded on the Stock Exchange Electronic Trading System ('SETS'), to meet the best execution requirement, firms should achieve a price (whether on SETS or an alternative execution venue) which at least matches the best price available on SETS)" (p. 13). The document continues, "However, in markets where there is no single or central exchange such as an over-the-counter ('OTC') market, the practice that is followed to substantiate the achievement of best execution is to obtain three quotes from different market makers (and then to select the most favourable price)" (p. 13).

The AIMR report was issued in November 2001. In its introduction, the report states that, "Therefore, it is not feasible to define a single measurement basis for best execution on a trade-by-trade basis. Instead, the guidelines focus on establishing processes, disclosures, and documentation, which together form a systematic, repeatable, and demonstrable approach to show that best execution compliance is consistent and effective."[14] The Investment Company Institute (ICI, an industry organization that represents institutional investors), in its comment on the AIMR Report puts this more strongly: "We recommend that the (AIMR) Guidelines clarify that best execution is not a quantifiable concept and that statistical measurements can be only one part of the overall assessment that firms may make in examining best execution."[15]

In assessing best execution, U.K. regulatory authorities give primary importance to the prices at which trades are made. Next in line is the timeliness of trades.[16] In Annex B of its Discussion Paper, the FSA presents brief summary statements about best execution from regulators in 13 different countries.[17] Five of the 13 summaries include explicit reference to the time dimension.[18] The regulators' focus on timeliness may be consistent with their focus on retail customers, because institutional participants commonly work their orders patiently in an attempt to get better prices.

---

[14] See AIMR [2001, op. cit.], p. 2.

[15] Letter to the AIMR. See A. Lancellotta, Letter to the Association for Investment Management and Research, Re: Proposed AIMR Trade Management Guidelines, *Investment Company Institute, Washington, DC* (February 12, 2002).

[16] Other considerations noted in the FSA's Discussion Paper include counterparty risk (see the discussion starting on page 26).

[17] The statements were extracted from the Forum of European Securities Commissions [2001].

[18] Phrases contained in the quoted references were "within a reasonable period of time," "as rapidly as possible," "without undue delay," "within a reasonable time period," and "as fast as possible."

For an institutional investor working a large order, the practice of patient order timing is not comprehended in a snapshot assessment of best execution. The FSA, as reflected in the following quote from its Discussion Paper, recognized this point in the context of its discussion of intra-day liquidity variations. It specifically noted that at certain times of the day, most notably at the open and at the close, spreads may be wide and that at such times it would be "advantageous not to deal" (p. 18). We would amend the statement to refer, more broadly, to price discovery, rather than to the bid-ask spread (which, of course, is also a relevant factor).

Our final comment about a regulatory authority imposing a best execution requirement is that the quality of executions achieved very much depends not just on the order placement decisions of individuals but on the efficiency of the marketplace that the individual participants are operating in. At the extreme, if a market becomes hyper-continuous under stressful conditions, best execution for an individual becomes a vacuous concept (by a hyper-continuous market, we meant that trades are occurring with great frequency, that transaction-to-transaction price volatility is highly accentuated, and that price discovery has broken down). Perhaps the regulators should place somewhat less emphasis on the handling of individual orders per se, and pay more attention to the quality of prices discovered for the broad market as orders are aggregated and turned into trades.

### The Elimination of Off-Board Trading Restrictions

Fixed commissions and off-board trading restrictions were established by an accord known as the *Buttonwood Tree Agreement* that was signed by 24 securities brokers on May 17, 1792 (the buttonwood tree was located at what is now 68 Wall Street). The New York Stock Exchange's restriction, NYSE Rule 390, had its origin in the Buttonwood Agreement. Rule 390 required that member firms take their orders for listed securities to the floor of an exchange for execution. Consolidation of the order flow was thereby forced, and competition between dealers and alternative market centers was limited.

The debate over off-board trading was far more heated, prolonged, and complex than the deliberations over the elimination of fixed commissions. As of 1975, NYSE Rule 394 prohibited both agency and principal executions away from the Exchange floor. After the Securities Acts Amendments of 1975, Rule 394 was modified (and its number changed to 390) to allow agency transactions away from the Exchange floor. Member firms, however, were still prohibited from participating as principals in off-board trading. Thus the Rule 390 prohibition continued to prevent the U.S. brokerage houses from assuming a dealership function, and thereby kept them from making their own markets in listed stocks.

Then, in 1980, member firms were allowed on an experimental basis to make off-board markets for all issues listed after April 26, 1979, and the SEC continued to press for the total elimination of the off-board trading restriction. Finally, the NYSE gave up and decided to rescind the rule. On May 5, 2000, the Commission approved the Exchange's request for a rule change and Rule 390 was gone.

But we are getting ahead of the story. The Exchange's off-board trading rule first came under scrutiny in a 1965 SEC confidential staff report. The SEC did not call for repeal of the rule, however, until the Commission was charged, by passage of the 1975 Amendments, "with an explicit and pervasive obligation to eliminate all competitive restraints that could not be justified by the purposes of the Exchange Act."[19] At the time, the NYSE was considered by many to be an exclusive club that only the privileged could join, and the Exchange's fixed commission structure had recently been found anticompetitive and unjustified. In this context, the off-board trading restrictions appeared to be another anticompetitive barrier that the Exchange had erected to protect its narrow self-interest. Consequently, in June 1977, the SEC announced that it would mandate the removal of Rule 390 by the end of the year.

In response, a cry of alarm was sounded by many market participants, certain academicians, and some members of Congress. The expressed fears over removal of the off-board trading restrictions carried far more conviction than the concerns voiced about the end of fixed commissions. It was apparent that the off-board trading issue was not as straightforward as the elimination of fixed commissions. In early December 1977, the SEC postponed the elimination of the rule.

What step might a regulatory authority take when there is no clear answer to a public policy issue? One possibility is to run an experiment, and this is what the SEC did. On July 1980, on an experimental basis, the Commission instituted Rule 19c.3, which freed Exchange members to make off-board markets for issues listed on the Exchange after April 26, 1979. The day before this rule became effective, the *Wall Street Journal* (July 17, 1980, p. 3) reported that Morgan Stanley and Merrill Lynch had both announced plans to trade in-house what had come to be known as *19c.3*

---

[19] Securities Acts Amendments of 1975, *Conference Report*. House of Representatives Report No. 94-229 (May 19, 1975), 94.

*stocks*. The article stated that "Most major broker-dealers say they will begin trading listed stocks if competitive pressures make such a move advisable." At the time, it appeared that substantial order flow might be diverted from the exchange markets.

By mid-1983, however, Merrill Lynch, Morgan Stanley, Goldman Sachs, and virtually all of the other large firms had stopped making markets in 19c.3 stocks. In-house market making is an inherently costly operation because of the considerable amount of financial capital that market making requires. The design of the 19c.3 experiment biased the results because in-house market making in the newly listed, 19c.3 stocks was not as profitable as market making in the established volume leaders such as IBM and Exxon might have been. And, in part, the stock exchange specialists simply had more fire power and so won the war: a brokerage house that competed with the specialists by making in-house markets for 19c.3 stocks might have received unfavorable treatment by the market makers when they turned to the exchange markets to rebalance their inventory positions. Whatever the reason, the experiment did not resolve very much, and the debate concerning the total removal of off-board trading restrictions continued.

Although the SEC's 19c.3 experiment failed to increase competition between market makers and did not resolve the consolidation/fragmentation debate, the Commission's involvement with the issue was successful in certain respects. To understand this, the following must be recognized:

- The Securities Acts Amendments of 1975 mandated the development of a national market system, but did not say precisely what such a system would be. Rather, the Act set forth certain broad goals to be achieved. One of the goals was fair competition among brokers, dealers, and markets. Another was the opportunity for investors' orders to meet without the participation of a dealer. These goals are vague and somewhat inconsistent. Understandably, the SEC had difficulty achieving them.
- The SEC had not been involved with the design of the market's architecture before the 1975 Amendments. Rather, the Commission had historically been responsible for setting regulations and procedures to prevent conduct contrary to the public interest and for overseeing and policing the markets. Accordingly, the new design responsibilities assigned to the SEC were not easily carried out by the Commission.
- The design and implementation of a national market system was impeded by the myriad conflicts of interest that prevailed among dealers, specialists, traders, exchange officials, issuing corporations, brokerage houses, self-regulatory groups, and the SEC itself. Thus the SEC was

responsible for achieving a vaguely formulated goal in the face of extensive opposition.[20]

- The legal power to mandate the removal of NYSE Rule 390 gave the Commission considerable control over the NYSE and various other groups in the industry. Whether by intention or by luck, the Commission's most successful move following the 1975 Amendments was to threaten, but not to put into effect, the total removal of the off-board trading restrictions.

Thus the Commission used its authority to remove Rule 390 to get the industry to move itself forward. The threat, made credible by the recent elimination of fixed commissions, worked. The exchanges opened their doors to new members, the clublike atmosphere of the NYSE has disappeared, and an intermarket linkage system was put in place. But NYSE Rule 390, though diminished, continued to stand.

The experience with off-board trading has been very different from the experience with fixed commissions. Rate fixing was widely viewed as anticompetitive and undesirable, but off-board trading restrictions appear to have some justification. Wherein lies the difference?

Here is the difference: As noted, there are externalities in trading, and there are public goods aspects to market making. Traders benefit from a market in which orders are consolidated, trading priority rules (e.g., price and time) are enforced, prices are set with reasonable accuracy, price movements are kept orderly, and trading is closely monitored. The superiority of such a market may not, however, guarantee its existence. The reason is that traders individually may have an incentive to turn to off-board markets. This can happen when network *externalities* are involved.

An off-board market can compete with a national exchange by guaranteeing timely executions at competitive, exchange-discovered prices. Because it does not assume the affirmative obligation to make a fair and orderly market, and because it does not provide the stock watch, market surveillance, and other services of an exchange, an off-board market may charge lower commissions than the exchange market. Assume this to be the case and that, consequently, an appreciable part of the order flow is diverted from a major market center. What might then ensue?

---

[20] In the Securities Acts Amendments, Congress had established the National Market Advisory Board (NMAB) to carry out the design function for the SEC. The board failed to achieve its objective within the two years it had been given. On December 12, 1977, after much wheel spinning and in the face of increasing impatience in Congress, the NMAB had its last meeting.

Price discovery on the exchange could become less accurate, and the investor protection provided by the exchange's stock watch and surveillance systems could deteriorate. In addition, spreads might increase and prices could become more volatile because of the reduced order flow. These changes would in turn cause the exchange to be a less-desirable trading arena. Accordingly, order flow to the exchange could decrease further, and the output of exchange services could be further impaired. In such an event, the quality of the off-board markets would also suffer. Consequently, the detrimental effects would be widely felt. Nevertheless, the prospect of this occurring would not be taken into account by investors when they make their individual trading decisions. Again we say that this is because of the network externalities and the public goods aspects of market making.

The NYSE market did not fragment soon after the rule was pulled. On the contrary, in 1999, the year before Rule 390 was removed, 82 percent of the share volume for NYSE listed securities was on the Exchange. In 2001, the first full year without the order consolidation rule, trading on the Big Board was an even more concentrated 84 percent of share volume.[21] However, as of October, 2008, the NYSE's market share stood at 28 percent: the electronic markets were making inroads into the NYSE market, the NYSE's own electronic trading platform was competing with its floor, and the NYSE floor activity was greatly diminished.[22]

A similar situation is developing in Europe. The initial pan-European regulatory initiative, The European Commission's *Investment Service Directive* (ISD) of 1993, empowered an EU member state, among other things, to institute a Rule 390-type order concentration rule that prescribes that orders be executed on a national stock exchange. Major European markets that had such a concentration rule included France, Italy and Spain. Effective November 1, 2007, the European Commission replaced the ISD with the landmark *Markets in Financial Instruments Directive* (MiFID). In so doing, the Commission has provided a regulatory regime for investment services for 30 member states in the European economic arena. Highlights of the legislation include a quest for enhanced intermarket competition; a call for greater transparency (most importantly for off-exchange trades); a requirement that firms take reasonable steps to obtain best execution (which the Directive recognizes includes, along with an execution price, the cost, speed, likelihood of execution, likelihood of settlement, and any other factor that may be deemed relevant); and, with regard to our discussion of NYSE Rule 390, MiFID has required the elimination of the order concentration rules.

---

[21] See NYSE 2001 Fact Book. The percentages refer to the percentage of volume reported on the Consolidated Tape in NYSE-listed issues. The participating market centers in addition to the NYSE were the Pacific, Chicago, Philadelphia, Boston, and Cincinnati stock exchanges; Instinet; and the National Association of Securities Dealers.

[22] The October 2008 market share statistic was provided by the NYSE.

Is the European regulatory initiative moving markets in a desirable direction? Within a year of MiFID's implementation, signs of fragmentation appeared in the European theater as market share for the national exchanges has dropped and new trading venues have emerged. *Ceteris paribus*, the intensified competition implied by the fragmentation bodes well for market quality items such as the level of commissions and other fees. How about aggregate liquidity provision, price volatility, and the sharpness of price discovery? For these market quality items, the answer is not so clear. It takes time for the players in a market to adjust fully to a regulatory change and for all of the consequences to be adequately judged. On this one, the jury is still out.

## Implicit Collusion

In 1994, two finance professors, William Christie and Paul Schultz, published a paper in *The Journal of Finance* that has had far-reaching consequences for the U.S. equity markets in general, and for broker/dealer firms and the NASDAQ Stock Market in particular.[23] At the time of their study, the minimum price variation (tick size) was one-eighth of a point, and Christie and Schultz had found that NASDAQ dealers were commonly avoiding the odd-eighth quotes (e.g., 60 5/8). We noted this occurrence in Chapter 6, pages 280 and 281. The authors suggested that dealers were "implicitly colluding" to keep spreads artificially wide. An investigation ensued of the NASDAQ market and the NASDAQ dealers by the Department of Justice and the SEC, and a class action lawsuit was filed. On November 9, 1998, approval of the court was issued for a settlement in the aggregate amount of $1,027 million.[24]

Further, communications between broker/dealers about market conditions were discouraged by a Department of Justice requirement that their conversations be taped. This has made price discovery more difficult in the NASDAQ market, particularly at market openings. Overall, the legal proceedings and the regulations that followed have had major impacts on how NASDAQ's dealer market operates.

---

[23] W. Christie and P. Schultz, "Why do NASDAQ Market Makers Avoid Odd-eighth Quotes?" *Journal of Finance* 49 (1994), 1813–1840.

[24] Opinion by the Honorable Robert W. Sweet, 94 Civ. 3996, U.S.D.C., Southern District of New York.

### The Order Handling Rules and Regulation ATS

Following the implicit collusion case, the SEC has continued to press for greater transparency of price and quote information. In 1997, the Commission instituted new Order Handling Rules which required any market maker firm holding a customer limit order to display that order in its quotes. Also, if a market maker has placed a more aggressively priced quote in an electronic market (a so called Electronic Communications Network, or ECN), its own NASDAQ quote must be improved to match its ECN quote. Concurrently, the Justice Department effectively stopped the NASDAQ market makers from communicating with each other by requiring that conversations be taped.

The Order Handling Rules let the proverbial cat out of the bag for they set the stage for new trading vehicles to enter the field. Before the rules were instituted, the broker-dealer firm Instinet was the one and only ECN. By September 1999 there were nine. To capture order flow, all a new ECN needed was to be a gateway that attracted some customers to place limit orders on its electronic book. Connectivity with other markets (either directly or through one of NASDAQ's own systems) would then enable market orders from the customers of other firms' to reach its book and trigger trades.

One of the new ECNs, Archipelago, stated on its web site that, "In January 1997, the U.S. Securities and Exchange Commission (SEC) implemented new Order Handling Rules that revolutionized trading in NASDAQ securities. The new rules created the opportunity for Electronic Communications Networks (ECNs), such as the Archipelago ECN, to interact directly with the NASDAQ National Market System. The Archipelago ECN was formed in December 1996 in response to these rules." Subsequently, in March 2006, Archipelago merged with the NYSE.

There is more. In 1998, an SEC release, *Regulation of Exchanges and Alternative Trading Systems* (known as Reg ATS), set forth rules that allowed an ATS the alternative of registering as either a national exchange or as a broker-dealer firm. Reg ATS has blurred the line between an exchange and a broker-dealer firm, has continued to promote a multiplicity of markets and, as a consequence, has further fragmented the order flow. The underlying reality is twofold:

1. A clear trade-off exists—intensifying intermarket competition (that is, competition in the market for markets) fragments a marketplace and, in so doing, can weaken competition between the buyers and sellers of shares (that is, competition within the order flow).
2. The very definition of an industry (including the breakdown between competitors and customers) can be profoundly affected by the regulatory framework that has been established by a government agency.

## Penny Pricing

In its efforts to further strengthen competition and reduce bid-ask spreads, the SEC pushed for a reduction in the tick size (minimum price variation), a topic we also touched on in Chapter 6. After much debate and discussion, the markets converted from fractional pricing to decimal pricing. The conversion went from an eighth to a sixteenth in 1997, and from sixteenths to pennies in 2001.

Standardizing the tick size across alternative market centers is good public policy for the following reason. Because the minimum price variation sets a floor on the size of bid-ask spreads (as we have discussed in Chapter 6), one market center can post quotes that are within the minimum spread of another market center if it quotes prices in smaller units. However, no market center wants to be forced to have noncompetitive quotes simply because of a tick size constraint and, consequently, no market center will choose a tick size that is larger than the smallest tick size of its competitors. But what is the smallest tick that can be selected? Ignoring the possibility of subpenny pricing, one cent is the smallest and, following the conversion in the United States to decimal pricing in 2001, all of the market centers started pricing in pennies. But does it make economic sense to quote prices in pennies? If a coarser unit of measurement would be preferable, the coarser tick size would have to be established by a regulatory authority to ensure the sustainable uniformity of a larger minimum price variation across the various competing markets.

Whether to mandate a switch from fractional to decimal pricing is not the critical question. The tricky public policy issue is how small (or large) the minimum price variation should be. We have seen that a relatively large price tick can keep spreads wider than they otherwise would be, especially for lower priced stocks. The wider spreads, of course, translate into higher returns per round trip for market makers and into higher transaction costs per round trip for customers. Does this justify lowering the tick size? The answer depends on whether or not the change has had other consequences. And it has.

The sharp reduction in the tick size from 12½ cents to 1 cent has narrowed spreads, but it has also made market making less profitable for the dealers. As a consequence, dealers firms (not surprisingly) have committed less capital to market making. And so spreads may indeed be tighter, but the number of shares posted at the best bid and offer has also decreased. What is the net impact on liquidity provision? The answer depends on the size of the customer who is seeking liquidity: the net impact may well be positive for a small retail trader who is looking to buy or to sell 1,000 shares or less, and negative for a large institutional participant who is looking to buy or to sell 100,000 shares or more.

Other consequences have followed the change to penny pricing. When one-eighth pricing was the norm, there were eight price-points per dollar. With pennies, there are 100. Consequently, less liquidity is concentrated at each price point, and depth at the inside market (the best bid and offer) has dropped off. As it has, the speed with which quotes change has accelerated. In the new regime, quote screens for actively traded stocks commonly flicker faster than the eye can follow.

A small minimum price variation affects participant trading strategies. An incoming order can gain priority in the trading queue by being placed just one cent above the best bid, or just one cent below the best offer. Being able to jump ahead of the queue for the price of a penny is called *pennying*. Pennying effectively undermines time priority, the secondary rule of order execution in an order driven market. Suppose that 10 orders have been placed at a best market bid of 32.00. If you were to enter your buy order at 32.00, you would be eleventh in the queue. However, if you were to place your buy order at 32.01, you would be first in the queue. In the old days with $\frac{1}{8}$ pricing, you would have to improve your price to 32.125 to get ahead of the queue at 32.00. Today you only have to go, as we have just said, to 32.01. And so, the cost of jumping ahead of a queue and gaining time priority has been reduced from 12.5¢ to 1¢.

But one person's gain of time priority is another person's loss. Once you have placed your order in the book, your time priority is established vis-à-vis other orders already existing at that price. But a newly arriving order can penny you and, in so doing, lower your standing in the queue of all orders that have been entered at your price and better. This loss of position in the queue can undermine your incentive (and the incentive of others) to place limit orders. In so doing, this discourages the provision of liquidity to the market. Quite clearly, rule making for a competitive equity market is no simple matter.

There was also talk about going to subpennies, but that did not happen. Andrew Brooks of T. Rowe Price had this to say: "Sub-pennies are illegitimate, except for a 5-cent stock. Maybe a 20-cent stock. But certainly not a thirty-dollar stock. Sub-pennies are confusing, they are distracting, and they bring no value to anybody. It is ridiculous for order flow to be directed to somebody because their bid or offer is better by a sub-penny."[25] With respect to penny pricing, William Christie stated that, "In retrospect, however, we were naïve in expecting pennies to balance the interests of the many trading constituencies, since a penny tick size has destroyed the critical roles played by price priority and limit orders. Any investor, including market specialists, can offer meaningless price improvement, step in front of existing limit orders, and render such orders virtually worthless."[26]

---

[25] See Robert A. Schwartz, John A. Byrne, and Antoinette Colaninno, eds. "Coping with Institutional Order Flow" (Springer Science+Business Media, 2005).

We conclude that fracturing the order flow over too many price points and discouraging the provision of dealer services and limit orders can be harmful to liquidity provision. The smaller tick size can result in higher intraday volatility and a less transparent market. As we have noted, following decimalization, the depth of orders at the inside market has decreased. And, not surprisingly, the dealers' commitment of capital to market making has fallen following this rule change.

## The Trade-Through Rule

When a transaction occurs at a price that is higher than the best-posted offer or lower than the best-posted bid and orders at these better prices are not included in the transaction, a *trade-through* is said to occur. Alternatively stated, the better-priced orders have been "traded through." For example, if a sell limit order on the book at 50.10 remains unexecuted after a trade takes place at 50.12, the 50.10 sell order has been traded through. Since 2007, trade-throughs have not been allowed in the U.S. equity markets. That is, strict price priority is enforced across all publicly displayed orders and markets. A key argument in favor of this rule is that it is fair to limit order traders, and that it encourages the placement of more limit orders. After all, the prospect of sitting with a limit sell order at 50.10 and seeing a trade being made at 50.12 does not do much to encourage a participant to place that limit sell at 50.10.

The rule is also seen to benefit market order traders because it assures that their orders will execute against the best available counterpart quote. Particularly with regard to market orders received by brokers from their retail customers, there was concern that, because of some other arrangements that a broker might have, the agent would not necessarily route an order to the execution venue that is posting the best price. This is a principal/agent problem that the Commission was right to be mindful of. However, a trade-through rule is not required to deal with this problem. A broker's best execution obligation covers it, and broker operations can be monitored using SEC reports that provide information on how the customer orders are being routed.

In 2005, the U.S. SEC in a major initiative implemented *Reg NMS* (short for Regulation National Market System). A centerpiece of Reg NMS is an order-protection rule that was designed to prevent the execution of trades at inferior prices. The regulation is commonly referred to as *the trade-through rule*. Because of the way in which the trade-through rule was specified, it effectively forced the New York Stock Exchange to institute an electronic trading platform, and to cut way back on its use of the trading floor. Specifically, Reg NMS says that fast, computerized marketplaces are not allowed to trade through each others' quotes, but that prices set by slower, floor-based markets can be traded through. Accordingly, to defend itself from being traded through, the Big Board had to institute a "fast" (that is, electronic) market.

---

[26] William G. Christie, "A Minimum Increment Solution," *Traders Magazine* (November 2003), 40.

Disallowing trade-throughs is equivalent to imposing price priorities across all markets that a stock trades in. Order-driven trading venues typically impose price priority rules for their own markets (and they typically have time priority rules as well). The thorny issue about the trade-through rule involves its application across different markets in a fragmented environment. The immediate problem is that price is not the only dimension of a trade, and an investor (particularly a large institutional customer) might have a good reason other than price for preferring one market venue over another.

Big players commonly slice and dice their orders, and submit the pieces to the market over a protracted period of time. Ultimately, the price paid or received for the entire package is what matters, not the price for each individual tranche per se. A large trader may be willing to "pay up" to execute a large order quickly before others see his or her footprints in the market and front-run the remaining parts of the order. If the individual pieces cannot be executed quickly, with certainty, under the institutional trader's control, and with anonymity, a worse price might be realized for the entire order. The bottom line is that a number of variables in addition to price matter to the big players, but the regulation focuses on the price variable only.

Reg NMS has indeed strengthened the intensity of competition among market centers, and it has prompted investments in the regional exchanges and various alternative trading systems. However, imposing strict price priorities across markets effectively prevents alternative trading venues from competing with each other in terms of one very important variable, the quality of price discovery that they can deliver. There is irony in this. The flip side of the U.S. regulatory policy that has promoted competition in the United States is market fragmentation. The regulators have thought that intermarket linkages along with the trade-through rule counter the undesirable effects of fragmentation while bolstering competition, but the trade-through rule itself undermines intermarket competition in one critical way: how effective is price discovery competition between market center competitors who are required by rule to arrive at the same price?

Another irony concerning the rule is that, while enforcing price priority across the different venues seemingly integrates different markets, the trade-through rule, in one way, can fragment the broader marketplace: A larger number of market venues can coexist because any one of the multiple trading facilities can receive order flow simply by posting a better price. That is, if an electronic market is posting a bid of 32.01 while the best-posted bid in all of the other fast markets is 32.00, an incoming market sell order must be sent to that market so as not to trade through its 32.01 bid. And so that market will get order flow, and the broad market will be fragmented. Yet, as we have already noted, price is not the only variable that matters; speed, a customer's control over his or her order, anonymity, and certainty of execution also count. But the trade through rule is monolithic—it looks only at price.

The removal of Rule 390 had previously freed NYSE member firms from sending their orders to an exchange, and the trade-through rule has provided additional impetus for this to happen. And, as we have noted, the rule effectively got the Big Board to institute an electronic trading platform, a defensive move that has drawn appreciable order flow away from its own trading floor. The bottom line? The 2007 SEC initiative was far from an innocuous rule change.

By the time the year 2008 had drawn to an end, the NYSE's market share had dropped to under 30 percent. Along with two national exchanges (the NYSE and NASDAQ) and five major regional exchanges, over 50 alternative trading systems are, at the time of this writing, operating in the U.S. marketplace. Does the trade-through rule along with technology (in the form of intermarket linkages) adequately integrate these markets, or is fragmentation in the current environment a serious problem? Has the competitive position of the NYSE been unduly weakened by regulatory intervention? Has the trade-through prohibition proven to be good public policy? The arguments both for and against the rule are multifaceted and, before it was instituted, the rule was hotly debated. It still can be. Clearly, articulating sound public policy with regard to an issue such as this is not a simple matter.

## CAVEATS FOR PUBLIC POLICY

This chapter, along with having set forth broad concepts concerning public policy issues, has paid substantial attention to a particularly complex micro market, an equity market. The health of the equity markets is critically important to a nation. Capital raising, economic prosperity, and growth depend on the ability of firms to obtain financial capital at reasonable cost, and the cost of equity capital depends in good part on market quality.

The imperfect workings (or worse the collapse) of a financial market under stress can have dire consequences for a national economy. For these reasons, issues concerning equity market operations have attracted major government attention in the United States and European arenas.

As long as there have been exchanges, participants have been concerned about market quality issues including honesty, transparency, fragmentation, liquidity, price volatility, and trading costs. Also of major importance are principal-agent issues, moral hazard problems, technological inertia, and a spectrum of abuses of power and position. Our overriding focus in this chapter has been on the role a government regulator should play with regard to the imperfections of the micro markets. Should the markets be left free to evolve naturally, or should a government agency participate in designing market structure and in writing the rulebook? This question is best addressed not in the abstract but with reference to a specific market that is facing a specific set of issues.

The rules of a market system must be set collectively by a body politic (a government and/or a self-regulatory organization), and further regulation may be called for if an unregulated market would fail to achieve a reasonably optimal distribution of resources at appropriate prices. As noted, the major causes of market failure for the equity markets include the market power of some traders, dealers, and market centers; externalities; the public goods aspects of marketability services; moral hazard problems, principal-agent relationships, and informational asymmetries; and technological inertia.

Following the enactment of the 1975 Securities Acts Amendments, government has played an increasingly active role in the evolution of the markets. First and foremost, the regulatory authorities and the industry's own self-regulatory organizations (SROs) have sought to ensure an honest market. In addition, in the United States, starting in the 1970s, Congress and the SEC have been concerned with the competitive efficiency of the market for broker/dealer services. Some attention has also been given (primarily by the SROs) to the accuracy of the share prices that are being set in the market centers. Price discovery is a difficult goal to comprehend and to implement, however, and it has received far less attention in the regulatory debates concerning the design of the trading systems.

Government involvement with market design has raised many questions. Market architecture is highly complex. There is a lot involved that students of the market do not understand or agree on. The very measurements of market quality and best execution are subject to considerable ambiguity. Market structure changes have unintended consequences, especially in a rapidly evolving technological environment. And when a government-mandated structural change goes awry, government recognition, understanding, and appropriate corrective action, if ever taken, may not be forthcoming for years. One could certainly argue that it is better to let alternative market centers make their own decisions, reap their own rewards, and accept the consequences of their mistakes as they battle for order flow in a competitive environment.

Is government intervention really required to ensure sufficient competition in the industry? On the one hand, we see technology innovation expanding the geographic reach of trading facilities and, increasingly, competition is taking place on a global scale. We also see the arrival of new firms and new trading modalities in the United States and Europe. On the other hand, major resistance to technological change persists, meaningful innovations are rare and far between, and the obstacles faced by a technology pioneer are daunting.

Because of network externalities, little is more important for the quality of a market than whether it receives sufficient order flow. Consequently, it is extraordinarily difficult for a new entrant to put competitive pressure on an already established market, especially if the new entrant offers independent price discovery. Regardless of its inherent efficiency, a newcomer simply may not get the order flow required for its potential to be realized. This being the case, the established market has a weakened incentive to innovate.

And so, the vibrancy of competition and the ineffectiveness of competition combine to make good public policy extremely difficult to formulate. For those who would like to see government play a more active role, we offer seven caveats.

1. Government agencies like to monitor what they regulate. This leads to much attention being given to performance characteristics that can easily be measured. Consequently, too much focus is on readily observable aspects of market quality (such as bid-ask spreads), and insufficient attention is directed to more amorphous matters (such as the accuracy of price discovery).
2. Antitrust regulation has a well-established history in the United States with regard to markets in general and the securities industry in particular. The U.S. regulatory authorities have been very concerned about the exercise of monopoly power, but there are counterarguments that should also be paid attention to:
   - Security markets have natural monopoly properties because of strong network externalities in trading.
   - Two markets are involved in security trading: the market for the shares of stock that are being traded, and the market for marketability services. And there is a tradeoff. Consolidating the order flow in a single market center intensifies competition between customer buy and sell orders for the traded shares, but it also weakens intermarket

competition. Reciprically, fragmenting the order flow across different market centers intensifies intermarket competition but weakens the competitive interaction within the order flow.

- Countering the power of a network is the ability of a satellite market to free-ride on the prices established in a major market center.
- Commissions and bid-ask spreads are only two components of transactions costs; market impact costs and errors in price discovery are likely to be of far greater importance, and these factors should not be ignored simply because they are difficult to measure. While monopoly power might keep commissions and spreads relatively high, there is little reason for it to keep market impact costs high and the quality of price discovery low. On the contrary, the concentration of order flow (read "liquidity") will dampen market impact costs and sharpen the accuracy of price discovery.

3. In 1999, eight leading exchanges in Europe harmonized certain key features of their market structures, and it is indeed desirable to synchronize various design features across different market centers. However, it is hazardous for a government agency to *mandate* structural change, especially across multiple markets. A mandated design feature is not easily withdrawn if it is found to be undesirable. A regulatory authority will commonly point elsewhere and try to fix the problem with further regulations. Once government involvement in market design starts, the process tends to become self-perpetuating.
4. With regard to fostering greater competition, considerable regulatory attention is commonly given to "fairness." Under the rubric of fairness, it is argued that participants should have equal access to markets, that the "playing field" should be level, that markets should be transparent, and so forth. Unfortunately, all participants are not equal, and free markets are not necessarily fair. In the United States, the SEC has sought, in the name of fairness, to protect retail customers. However, the point has been widely made that an institutional investor, be it a mutual fund, pension fund, or other, is itself little more than an amalgam of many small individuals. Furthermore, all too often, a firm that is being threatened by competitive pressure uses the "fairness" argument. Rather than strengthening its own efficiency, the beleaguered player seeks protection through regulatory intervention.
5. Vested interests and technological inertia exist. To these, we add a third: The regulatory process itself can stymie innovation. In a competitive environment, innovations need to be made in a timely manner, but obtaining necessary regulatory approvals is typically a lengthy process. Furthermore, the power to deny change conveys the power to dictate change.
6. A marketplace in certain respects operates as an ecology, and regulation that affects one facet of market operations can have unintended

consequences on other parts of the "ecological system." This introduces a further element of risk into regulatory intervention, especially if the economic dynamics of the market are not adequately understood by the regulatory authority.

As time rolls on, micro markets will continue to evolve in response to the interacting forces of competition, technology, and regulation. As they do, the public policy issues that they give rise to will shift and evolve as well. In order for these issues to be dealt with appropriately, it is essential that the strengths, shortcomings, and interactions between technology change, government intervention, and free market competition be properly understood and respected. To this end, we emphasize the following thoughts:

- Regarding technology, it is important to keep in mind that it is neutral, that is, that technology's contribution to aggregate welfare depends on the uses to which it is put.
- Regarding government intervention, the public policy issue of deregulation versus reregulation should not be approached solely as an abstract, philosophical question. Appropriate regulation is what is needed, the objective is not at all easy to achieve, and each case will likely require its own distinctive solutions. It is important to keep in mind the maxim that we have cited, that "there should be as much regulation as necessary, yet as little as is possible." Also recall Montesquieu's dictum that "Useless laws weaken necessary laws."
- Regarding competition, it is indeed the case that the free market does not always deliver the results that we want (nothing is ever perfect, not even close), and the marketplace can be a very harsh master.

  But harshness has a positive side. As we noted in Chapter 7 (in the Section, Markets at Risk: the Financial Turmoil of 2007 to 2009), an irrevocable interplay exists between the forces of growth and the forces of destruction. We see it in Schumpeter's concept of creative destruction. We sense it in Adam Smith's concept of the invisible hand. But economists do not have a monopoly on this thought, for it is also apparent in attributes that the poet Percy Bysshe Shelley (1792–1822) has attributed to the west wind: "Wild Spirit, which art moving everywhere; Destroyer and preserver."

7. Our seventh caveat is that the strengths of the micro markets be honored, that we not lose faith in the efficacy of a free market.

## CHAPTER SUMMARY

Public policy issues pertaining to the operations of micro markets evolve as the markets respond to three powerful forces: competition, technology, and regulation. We focused on competition in Chapter 6, and in Chapter 7 we considered conditions under which competitive micro markets will fail to deliver socially desirable results. Government regulation, a topic that naturally follows from the material in Chapter 7, is the number one topic for this chapter. In dealing with it, we first considered three integrally related issues: governance structure, a market's self-regulatory obligations, and technology. Throughout, we paid particular attention to securities markets. These markets are complex, they are of great importance to the broad economy, and they provide a rich opportunity to address a spectrum of public policy issues in a relatively concrete fashion rather than in more abstract terms. Here are the highlights.

- Regulation of the markets has four major targets. First and foremost is the enforcement of honest behavior. Simply put, trust is of overriding importance; without it, market quality is at best undermined and, at worst, a market may actually collapse. The second target is the quality of free market outcomes, because the prices and quantities established in an unregulated market may be deemed undesirable from a public policy perspective. Macroeconomic policy is the third target, with monetary and fiscal policy being the primary macroeconomic tools. Fourth is the regulation of market structure. For over three decades in the United States, the structure of the equity markets has been subjected to government regulation and, as has been our way throughout this book, it is to these markets that we have directed our attention.
- One of the first things to understand about a micro market is its governance structure. In recent years, many securities exchanges around the globe have changed from not-for-profit membership organizations to being privatized, for-profit companies. In reviewing this development, we considered the customers of an exchange (Are they its members or the public traders?), the operations of an exchange (Does it provide a service to its members or does it run a trading platform for public participants?), the dynamic development of exchange facilities (Is growth efficient, or is innovation being stymied by technological inertia?), the provision of public goods–type services (including price discovery, liquidity creation, and the tempering of accentuated intraday price volatility), and the branding of a market such as the New York Stock Exchange (as provided by its trading floor and self-regulatory operations).
- Stock exchanges are typically among the most regulated of the micro markets. They operate within the context of basic rules, common practices, and regulation that is self-established and/or externally imposed. When it sets, supervises, and enforces its own rules, an exchange is a Self-Regulatory Organization (SRO). The overall objectives of

exchange rules and regulations are investor protection, system protection, and the achievement of a "fair and orderly" market.

- The delegation to an exchange of the power to set its own rules and regs is usually stated in a national or federal law. When an exchange establishes its own rules, approval by a national or federal securities commission is usually required. Ideally, government regulation and self-regulation should complement rather than conflict with each other.
- Involving people who are closest to the action in rule-making has advantages in terms of knowledge and speed (the rapidity with which decisions can be made and implemented), and self-regulation facilitates realizing these benefits. The negatives include the possibility of competition being impaired by cartelized practices, by weaker players being placed at an unfair disadvantage, and by regulatory capture that happens when stronger players sway the regulatory process.
- For the past several decades, technology development has transformed many micro markets in general and the equity markets in particular. We reviewed some of these developments, paying particular attention to the NASDAQ Stock Market in the United States, a market that has pioneered the introduction of electronic technology and that has totally reengineered itself by its technology innovations.
- We noted the far-reaching consequences that technology development and automation have had for employment opportunities, consumption decisions, productivity, economic growth, and the quality of life. While many of the benefits are self-evident, we stressed that technology itself is neutral, that technology is a tool that enables tasks to be accomplished, and that technology's social desirability depends on the tasks to which it is put and on how the tool is used.
- Automation has had an extensive impact throughout the securities markets:
  - The rapid calculation of price indices has supported the trading of new financial products (e.g., index futures and options), and has provided valuable pricing guides for individual shares.
  - Automated trading provides for a faster, more error-free transmission of trade data into posttrade clearance and settlement, has facilitated the overview and regulation of trading, and has been a boon for industry analysts and academic research.
  - The speed with which events happen in an electronic marketplace (along with the practice of "slicing and dicing") have led to the use of automated order writing and submission. The procedure, known as *algorithmic trading*, is not just a product of automation; it is also a symptom of the complexities that high speed electronic trading can introduce.

- Automation has posed other challenges for equity trading. Unresolved issues persist concerning the electronic handling of large, institutional-sized orders, and the computerized trading of relatively illiquid mid- and small-cap stocks. Trading itself is a challenging application to automate, as the software that makes an electronic market operational has to be specified in far greater detail than is required for human-intermediated trading. Moreover, software has to be rewritten to implement any new market structure change (market structure can evolve a good deal more flexibly in a human-intermediated environment).
- As technology change transforms economic processes and in so doing alters the distribution of wealth and power across economic agents, regulatory intervention must also be reshaped to keep up with the evolving dynamics of the micro markets.
- The chapter has dealt with the evolving focus of U.S. government regulators, going back to the Congressional Securities Acts Amendments of 1975 (which precluded any return to fixed commissions and mandated the development of a National Market System). In the years that followed, extensive regulatory change has been experienced on both sides of the Atlantic, as government has become far more involved in the structure of the equity markets.
- We considered in further detail three major regulatory goals:
  - Assure an honest market.
  - Assure an efficient provision and pricing of broker/dealer services.
  - Assure the efficient and reasonably accurate pricing of the shares that are traded.
- In parallel with our discussion in Chapter 7 of the efficiency (or lack thereof) of free market competition, we next considered a number of ways in which government regulatory intervention can itself be flawed. Here is the list of items we discussed:
  - Inadequate procedural safeguards.
  - Existence of jurisdictional disputes.
  - Cost to society of regulating markets.
  - Cost to the markets of being regulated.
  - Incomplete and imperfect knowledge on the part of the regulators.
- We then switched from an aerial view of the problems concerning government intervention and took a closer look at how the U.S. Congress and the U.S. Securities and Exchange Commission (SEC) have intervened in the structure and operations of the U.S. securities markets. The following seven regulatory initiatives were discussed:
  - The end of fixed commissions as mandated by the Congressional Securities Acts Amendments of 1975.

  - The imposition of a best-execution obligation (also part of the 1975 Securities Acts Amendments).
  - The elimination of off-board trading restrictions (which, in the United States, was accomplished in 2000).
  - The implicit collusion case against NASDAQ dealers that had its origin in Christie and Schultz's 1994 landmark paper in the *Journal of Finance*.
  - The Order Handling Rules and Regulation ATS.
  - The change to penny pricing in 2001.
  - A 2005 SEC release referred to as Reg NMS that established an order protection rule that was designed to prevent the execution of trades at inferior prices (aka the trade-through rule).
- In the section of the chapter, "Caveats for Public Policy," we raised the question of whether markets, with all of their imperfections, should be left free to evolve naturally, or whether a government agency should participate in designing market structure and in writing the rulebook. Once again we addressed the question not in the abstract but with reference to a specific market, the equities market.

  We said: "Market architecture is highly complex. There is a lot involved that students of the market do not understand or agree on. The very measurements of market quality and best execution are subject to considerable ambiguity. Market structure changes have unintended consequences, especially in a rapidly evolving technological environment. And when a government-mandated structural change goes awry, government recognition, understanding, and appropriate corrective action, if ever taken, may not be forthcoming for years."
- We went on to say, "The vibrancy of competition and the ineffectiveness of competition combine to make good public policy extremely difficult to formulate. For those who would like to see government play a more active role, we offer the following caveats":
  - Government agencies, liking to monitor what they regulate, tend to focus on readily observable aspects of market quality such as bid-ask spreads, and they pay insufficient attention to more amorphous matters such as the accuracy of price discovery.
  - While regulatory authorities are very focused on the exercise of monopoly power, there are other considerations that should be paid attention to: network externalities, free-riding on price discovery, and the nonexplicit components of trading costs such as market impact costs and errors in price discovery.
  - Once government becomes involved in market design, the process tends to become self-perpetuating.
  - Considerable regulatory attention is commonly given to "fairness," to "leveling the playing field." But all participants are not equal, free

  markets are not necessarily fair, and all too often a firm that is threatened by competitive pressure uses the fairness argument in an appeal for regulatory intervention rather than itself strengthening its position in the market.
  - Vested interests and technological inertia certainly exist. But the regulatory process itself can also stymie innovation, in part because obtaining necessary regulatory approvals is typically a lengthy and difficult process.
  - A marketplace in certain respects is an ecology, and regulation that addresses one facet of a market's operations can have unintended consequences on other parts of the ecological system.
- The free market does not always effectively deliver the results that we may desire from a public policy point of view, and the free market can be a very harsh master. But nothing is ever close to being perfect, and that includes both the markets themselves and the governments that might regulate them. The public policy issue of deregulation versus reregulation should not be approached solely as an abstract question. Appropriate regulation is certainly needed, but achieving it is not at all easy.
- We quoted the poet Shelley concerning the power of the west wind both to destroy and to preserve, and ended the chapter with this thought: "Our seventh caveat is that the strengths of the micro markets be honored, that we do not lose faith in the efficacy of a free market."

## QUESTIONS TO THINK ABOUT

1. We started the chapter with the statement that markets around the world are being shaped by three powerful forces: competition, technology, and regulation. For much of the chapter we have focused on the third. From a public policy perspective, appropriate government regulation is certainly needed for the micro markets to operate effectively, and yet, unless properly structured, regulation can unduly impair the workings of a free market. Quite clearly, achieving an optimal balance between the competitive forces of a marketplace and the requirements that are imposed by a regulator is far from a simple matter. What are your thoughts about this issue? Has your thinking about it been affected by your study of microeconomics? Please explain.
2. Market efficiency was our topic in Chapter 7, and in dealing with it we considered the financial turmoil that first emerged in 2007

and which erupted with great force in the United States and world markets in fall 2008. What are your thoughts about the broad causes of the market breakdown and your ideas concerning steps that a government should take to rectify the situation? Has your thinking about this issue been affected by the discussion in Chapter 8? Please explain.

3. In your opinion, to what extent did the economic picture in the first decade of the twenty-first century resemble the wild and woolly scene that characterized the markets almost three-quarters of a century earlier? What have we learned from these experiences, and what do you think should be done about all of this from a public policy perspective? While thinking about these questions, consider the following:
    - As a major deregulatory step, Glass-Steagall was repealed in 1999.
    - The collapse of the financial markets that started in fall 2008 and the broader recession that followed occurred at a time characterized by the emergence of new, complex, and poorly understood (by many) financial instruments (credit default swaps, for instance).
    - In recent years, relatively nontransparent alternatives to traditional banking have emerged. These include hedge funds, private equity firms, venture capital firms, and so on. Their operations, as of this writing, have been largely outside the scope of the regulatory authorities.
    - Over the years, banking regulation has remained fragmented across an array of state banking commissions, the Federal Reserve System, and the Treasury Department.
    - Major scandals have erupted (Enron's collapse in 2001, and the Bernard L. Madoff scam which came to light in 2008, to name two).
    - Greed was rife. As the economy was spiraling down, much public attention was directed to the outsized salaries and bonuses that were being paid out in the financial services industry. A huge public outcry attended news of the bonuses (which were also referred to as contractual retention agreements) that the teetering insurance giant, AIG, paid out at a time (2009) when it was receiving federal bailout money.

(*Continued*)

(*Continued*)

4. Financial services are vitally important for micro markets, and executives in financial services firms can realize extraordinary wealth from monies obtained through salaries, bonuses, stock options, and so on. Consider another resource, water. Water is of overriding importance for all of life, and water is generally not free. Yet corporate incomes realized from the sale of water are nowhere near what they are in the financial services industry. What accounts for the difference? The complexity of the product and the specialized ability of many of the people in the financial services industry are obviously a big part of the answer. But how exactly does this work? As we have seen, microeconomic analysis shows that in a competitive environment the price of a good or service should be in line with its marginal cost of production. Is this condition more apt to hold for water than for the provision of financial services? What do you think? How good a job does a free market do in pricing highly special resources such as top-of-the-line corporate executives, highly skilled (and specialized) employees, superstar athletes, and one-of-a-kind Picasso paintings? What do you think? Do you believe that there is a role for government to play with regard to any of this?
5. We ended this chapter by quoting the poet Shelley, who described the west wind as a destroyer and preserver. Shelley was thinking of the seasonal cycle where the west wind blows away the dead leaves of fall so that new growth can emerge in the spring. What is the relationship between Shelley's west wind, Smith's invisible hand, and Schumpeter's concept of creative destruction? Might a seasonal cycle and a business cycle have something in common? What are your thoughts about all of this?
6. Select a nonfinancial micro market that you have some familiarity with and that has been subjected to meaningful government regulation. Describe the regulation and the economic justification for having imposed it. Do any of the causes of market failure that we enumerated in Chapter 7 justify the regulation? What is your opinion about the success of the regulation?
7. We focused in this chapter on seven major regulatory initiatives that the U.S. government has taken with regard to the U.S. equity markets, starting with the Securities Acts Amendments of 1975. While the first initiative, the preclusion of fixed commissions, has been broadly acclaimed as a success, we did not come to firm

conclusions about the others. Instead, we noted some points of justification, some points of concern, and stressed the complexity of the issues. What is your thinking about the initiatives? You might approach this question by looking at each regulation individually, and/or by thinking more broadly about the direction in which they have collectively taken our markets.

8. For equity trading, market quality has to do with the markets being reasonably transparent (can participants get adequate and timely information about market quotes and trades?), being reasonably consolidated (is the marketplace concentrated to an extent that raises fear of monopoly power, or is it fragmented to an extent that impairs reasonable price and quantity discovery?), and being reasonably efficient in terms of price discovery (do prices quickly adjust to new equilibrium values following any informational shock?) What is your assessment of the U.S. government's regulatory initiatives in light of these market quality goals?
9. The U.S. SEC's regulatory framework is *rules-based*, which means that the agency typically prescribes detailed, specific rules concerning the responsibilities of its constituents. An alternative framework is a *principles-based* approach, which means that the regulating authority states broad principles for compliance, and offers the regulated parties some flexibility in applying them to their specific circumstances. The principles-based approach is followed by the Financial Services Authority (FSA) in the United Kingdom. There is a trade-off between the two alternatives: The rules-based approach is more prescribed and therefore less ambiguous, while the principles-based approach offers greater flexibility. Which approach do you think is preferable?
10. Philanthropy was the topic of our end-of-chapter application for Chapter 2. In the context of the current chapter, we can think of a household's or a firm's charitable giving as a private effort to improve social welfare (this view certainly applies to private giving to, for instance, universities, hospitals, and the arts). In this context, think about the implications that philanthropic giving can have for the following:
    - The desirability from a public policy perspective of much charitable giving being tax deductible.
    - The role of government versus the role of private citizens in determining the allocation of funds to social (public) projects.

(*Continued*)

(*Continued*)

- The drive of some private citizens to increase their income and wealth past a point that others might characterize as "greed."
- With regard to most goods and services that we consume, it is likely that, beyond a certain consumption level, we reach satiation. The utility function that we dealt with in Chapter 2, however, is not thought of as having a finite peak. What mindset on the part of a utility-maximizing decision maker might explain the absence of a finite maximum?

## TraderEx SIMULATION: ECONOMIC EFFECTS OF AN ORDER PROTECTION RULE

Once again, we turn to the TraderEx simulation software that is available in *Mastering the Art of Equity Trading Through Simulation: The TraderEx Course*. In this chapter we have looked at a number of regulatory initiatives that the U.S. Congress and the U.S. Securities and Exchange Commission have taken since 1975 to strengthen the U.S. equity markets. Regulations of this nature can have far-ranging impacts, unintended consequences can occur, and mistakes may prove to be very costly. Thus the Commission has proceeded with care, seeking much advice and council, including formal requests for comments from the public.

In the simulation exercise for this chapter, we focus on one of the SEC initiatives that we have discussed, the trade-through rule, which was instituted as part of the SEC's 2005 Reg NMS release. Our purpose is to give you further insight into how the rule can affect the placement of orders in an order-driven equity market. Here is the setting:

1. Two trading venues are used in this exercise: (1) a transparent limit order book, and (2) a block trading facility.
2. Each simulation is run for one trading day (roughly 10 minutes of clock time).
3. Your task for each run is to execute a 3,000-unit buy order. Your success will be assessed by: (1) how completely you have filled your order (did you buy all 3,000 units?) and (2) your average buying price relative to VWAP (did you beat the stock's volume weighted average price over the course of the simulation run?).

4. Run your simulations for each of three settings:
   A. Limit order book only.
   B. Limit order book and block trading facility with no trade-through rule.
   C. Limit order book and block trading facility with the trade-through rule enforced.
5. For each of the three settings, run the simulation a few times so that you get an idea of the distributional properties of your results.
6. Keep count of the number of trades that you have made in each simulation run and of the number of limit orders that you have placed.
7. Additional details for the simulation exercise are provided in the TraderEx book.

Please respond to the following:

1. Contrast setting A with setting B. How did the inclusion of the block trading facility affect your order placement? How did it affect your two performance measures?
2. Contrast settings B and C. How did the inclusion of a trade-through rule affect your order placement strategy? How did it affect your two performance measures?
3. Did any trade-throughs occur when the trade-through rule was not enforced? If so, how do you feel about it?
4. What are your thoughts now about (1) the desirability of having the block trading facility, and (2) the desirability of having the trade-through rule enforced?
5. Run the simulation again for settings B and C with your task being to fill a 5,000-unit order. How might increasing your order size affect your thoughts about the desirability of instituting a trade-through rule?

# Glossary

## A

**adaptive valuations** As applied to equity trading, the act of an investor changing from a bearish valuation to a bullish valuation, or vice versa, upon learning the valuations of others. A seller becoming a buyer, or vice versa.

**administered prices** Prices that are set by firms that have the market power to do so.

**algorithmic trading** Computer driven trading. Algos are computer-resident trading rules that electronically generate buy and sell orders.

**allocational efficiency** Producing and consuming goods and services in optimal proportions. For instance, a consumer is optimally allocated income between two goods (call them *X* and *Y*) when the ratio of marginal utilities for the two goods equals the ratio of their market prices. Similarly, is efficiently using two inputs into production (call them labor and capital) when the ratio of their marginal products equals the ratio of their market prices.

**arbitrage trading** The coordinated trading of two or more assets that are mispriced vis-à-vis each other. For instance, the purchase of a relatively low-priced stock and the simultaneous short sale of a relatively high-priced stock. The arbitrageur profits as the prices of the two stocks converge (as the spread between them shrinks).

**arbitrageur ("arb")** Somebody who engages in arbitrage trading.

**ask quotation** The price at which someone is willing to sell shares.

**asymmetric information** Some market participants are in possession of information that other participants do not have.

**asynchronous order flow** The staggered sequence in which buy and sell orders arrive at the market. Alternatively stated, buy and sell orders do not arrive at the market at the same time.

**average fixed costs** A firm's total fixed costs divided by total output.

**average variable costs** A firm's total variable costs divided by total output.

## B

**barriers to entry** Anything that precludes a new seller from entering a market. For example, a government license, patent protection, control over a scarce resource, and substantial economies of scale are barriers.

**behavioral economics** Analysis of economic decisions that are made by agents, not entirely according to rational economic principles, but also in light of cognitive and emotional factors, and bounded rationality.

**bid quotation** The market posted price at which someone is willing to buy shares.

**bid-ask spread** The gap between the lowest quoted offer that has been posted on the market and the highest quoted bid that is posted on the market.

**bookbuilding** We use the term in this book to describe the process of buy orders and sell orders being placed on a limit order book so that the book will become deeper (offer more liquidity). Bookbuilding in the secondary equity markets is particularly important at the start of each trading day. The term is more commonly used in relation to building investor interest in a stock at the time of its initial public offering (IPO).

**brokers** Sell-side participants who handle orders for their customers as agents. Brokers typically receive commission payments for their services.

**budget constraint** The constraint that identifies the alternative combinations of goods and services that a decision maker can obtain given prices and the decision maker's income ($M$). In a two-good model ($X$ and $Y$), the constraint is the linear equation,

$$Y = \frac{M}{P_Y} - \left(\frac{P_X}{P_Y}\right)X$$

**buy-side customers** Investors who require the services of sell-side intermediaries to get their trades done. Institutional investors such as mutual funds, pension funds, and hedge funds are buy-side customers.

## C

**call auction** An order driven market that provides periodic rather than continuous trading. All submitted orders are cleared at the same point in time at the same price in one big, generally multilateral trade. Electronic calls are commonly being used to open and close a trading session.

**capitalization (cap) size** The total value of shares (the price per share times the number of shares) of a stock that is outstanding.

**capital (types of)** Physical capital (such as plant and equipment), human capital (such as education and health), and financial capital (such as equities, bonds, and derivatives), and intangible capital (goodwill).

**Capital Asset Pricing Model** (CAPM) A cornerstone of modern portfolio theory that shows how risk and expected return are related in the market, and how individual stocks are priced given their systematic risk (their covariance with the market portfolio).

**capital market line** An upward sloping line that delineates the risk-return trade-off that investors face in a CAPM framework when combining shares of a market portfolio with a risk-free asset. The capital market line plays the same role with respect to the risk-return trade-off as the standard budget constraint plays with respect to the generic $X$, $Y$ trade-off.

***ceteris paribus*** Latin for "other things being equal." A methodological treatment that enables the net effect of one variable (an independent variable) on another variable (a dependent variable) to be obtained, while holding all other relevant independent variables constant.

**clearing price** The price set in call auction trading at the time when the market is called and the price is set. The clearing price is typically the value that would maximize the number of shares that trade. Equivalently, the clearing price is the value that best balances the cumulated buy and sell orders that have been submitted to the call. Buy orders that are equal to or higher than the clearing price execute, as do sell orders that are equal to or lower than the clearing price.

**commission bundling** The provision of additional services (e.g., research reports, data sets, or tickets to a choice sporting event) by brokerage houses that commission revenue pays for.

**comparative statics** A contrast of different equilibria for a dependent variable for different values of an independent variable, all else constant.

**complements** Products that are consumed in combination with each other. If $X$ and $Y$ are complements, the cross elasticity of demand for $X$ with respect to the price of $Y$ is negative. Tennis rackets and tennis balls are complements. With perfect complements, $X$ and $Y$ are consumed in fixed proportions regardless of their relative prices.

**consumer surplus** The (larger) total amount that a purchaser (consumer) would be willing to pay for $X$ units of a good (the purchaser's reservation price times $X$) less the (smaller) total amount that has to be paid (the market price for $X$ times $X$) because the market price of $X$ is less than the purchaser's reservation price for $X$.

**contestable markets** Markets that face the threat of competition from potential new entrants.

**continuous trading** An order driven or quote driven market where a trade can be made at any point in continuous time that the price a buyer is willing to pay (or that a seller is willing to receive) crosses a price that someone else is willing to receive (or is willing to pay).

**contract curve** A line showing the efficient allocations of two goods between two participants. The locus of all points of tangency between the indifference curves of the two participants when the indifference curves of one of them have been rotated 180 degrees. See "Edgeworth box."

**cost of ignorance** The cost to a dealer firm or limit-order placer of trading when dealing with a better-informed participant.

**counterparty risk** The risk that the contraparty to a contract will fail to perform its obligations under that contract. In equity trading, the risk that the contraparty to a trade will not deliver shares (to a buyer) or cash (to a seller).

**covariance with the market** A measure of how the rate of return for an asset varies with the rate of return on the market portfolio.

**cross-price elasticity** The percentage change in the quantity demanded or supplied of a good or service divided by the percentage change in another good or service's price, all else constant.

## D

**dark liquidity pools** Trading facilities where the buy and sell orders of participants are not disclosed. Opaque markets.

**deadweight loss** The reduction of total economic surplus (consumer surplus plus producer surplus) when output is less than the frictionless, free market equilibrium output, and price is higher for buyers and/or lower for sellers.

**dealers** Sell-side participants who trade with their customers and, in so doing, take principal positions. Dealers typical receive their compensation from the bid-ask spread.

**demand curve** A graph of the relationship between the quantity demanded of a good and the price of the good, all else constant.

**divergent expectations** Different investors in possession of the same information have different risk and return expectations for a security. Heterogeneous as distinct from homogeneous expectations.

**diversifiable risk** That portion of the total returns variance for an asset that can be eliminated by portfolio diversification.

**duopoly** A market in which there are just two firms competing with each other.

**dynamic dealer pricing** Adjustment made in a dealer's bid and ask quotes to control inventory imbalances.

## E

**economic information** Information concerning the firm's product market, the firm's competitors, national and international economic conditions, and so on.

**economic rent** Excess returns ($TR - TC > 0$) to a firm that cannot be competed away in the long run by other firms entering the industry. Returns (for a firm) that are attributable to a special advantage that the firm enjoys (such as may be offered by its location). Rents are differentiated from economic profits, which can be competed away in the long run.

**economies of scale** Economies achieved by a firm from producing a larger output. The relationship between long-run average cost and output. A firm's long-run average cost curve is decreasing over the range of output for which it is experiencing economies of scale. Long-run average costs are constant under constant returns to scale, and are rising under diseconomies of scale.

**economies of scope** Economies achieved by a firm from the production and distribution of multiple products. Economies of scope are differentiated from economies of scale.

**Edgeworth box** A diagram showing all possible allocations of given amounts of two goods between two consumers. See "contract curve."

**effective demand** The amount of a good that a buyer would actually be willing to purchase.

**efficient market hypothesis (EMH)** The hypothesis that all existing information is fully reflected in share values. No trader can trade profitably (realize above-normal risk-adjusted returns) based on already-existing information.

**elastic price demand** A percentage change in price induces a greater percentage change in the quantity demanded.

**elasticity** The responsiveness of demand or supply to one of its determinants. For instance, let *Y* be a dependent variable and *X* be an independent variable. The elasticity of *Y* with respect to *X* is the percentage change in *Y* divided by the percentage change in *X*. Perhaps *X* is the quantity of a good demanded and *Y* is the unit price of *X* (own price elasticity). Or *Y* could be income (income elasticity), or *Y* could be the price of another good that is related in consumption (cross-price elasticity).

**equities** The indicia of ownership in an economic enterprise; typically common stock.

**equity market microstructure** The analysis of the detailed way in which orders to trade equity shares are submitted to a market, handled by the market, and turned into trades and transaction prices. Of particular importance is how the design and operation of a specific market permits reasonably accurate price discovery and reasonably complete quantity discovery.

**exchange traded funds** (ETFs) An investment vehicle that is traded on an exchange much like ordinary shares. An ETF holds stocks, bonds, or other assets in trust, and shares of the ETF trade at roughly the asset value of the underlying assets that comprise the ETF.

**execution costs** The implicit costs of trading: for instance, bid-ask spreads and market impact.

**expansion path** With the quantity of labor and the quantity of capital inputs on the axes, the expansion path is the line established by the points of tangency between the isoquants and the isocost curves.

**expected return** The return that a risky stock or stock portfolio is expected to yield as of a future period of time. If returns are normally (lognormally) distributed, the expected return is the mean of the normal (lognormal) distribution.

**expected utility** The utility that a decision maker expects to achieve from acquiring a risky position.

**explicit costs** The direct costs of trading: for instance, commissions and taxes.

## F

**factor of production** An element (but not an ingredient as flour is for cake) in production for which there is supply, demand, and a market determined price. Productive factors are traditionally classified as land, labor, and capital as identified by Adam Smith. We have also included information and financial capital as factors of production. Closely associated with financial capital is liquidity, a more encompassing but more abstract concept.

**fails to deliver** The failure of a seller to deliver securities or cash within a settlement period (usually two days).

**fair bet** A risk situation with an expected monetary return of zero.

**financial information** Information concerning a firm's current capital structure, earnings forecasts, and so on.

**fixed cost** Costs of production that do not vary with output. Fixed costs exist in the short run. In the long run all costs are variable.

**flow variable** A variable that is measured as a rate per unit of time (in contrast with a stock dimensioned variable that is measured as of a point in time). For instance, the amount of a resource consumed (or output produced) during an hour, a day, a month, or a year. Income, consumption, and stock returns are examples of flow-dimensioned variables.

**friction and frictionless** These opposites reflect the realities of market frictions (the costs, blockages, and imperfect information) compared to a theoretical frictionless market without costs and blockages and characterized by perfect information.

**fundamental information** Economic and financial information pertaining to the determinants of future share values; *fundamental* as distinct from *market* information.

**fungible resources** Resources (commodities) that are identical (indistinguishable) one from the other. Fungible resources are perfect substitutes, one for the other.

## H

**high-frequency data** Intraday trade and quotes data. The complete record of all trades and quotes.

**homogeneous expectations** Investors in possession of the same information form the same risk and return expectations for a security; *homogeneous* as distinct from *divergent* (i.e., *heterogeneous*) expectations.

## I

**imperfect competition** Characterizes a market that is neither perfectly competitive nor monopolized; monopolistic competition and oligopolistic competition.

**income elasticity of demand** The percentage change in quantity for a good or service divided by the percentage change in a consumer's income, all else constant.

**income expansion path** For a two-good model (e.g., *X* and *Y*), the locus of optimality points in X,Y space achieved by a consumer for different values of the consumer's income, the prices of X and of Y given (i.e., constant). For a consumer, the locus of all tangency points between a family of indifference curves and a set of parallel budget constraints that represent different levels of income.

**indicated clearing price** The call auction price that is displayed to the public while orders are being entered into the auction during its bookbuilding phase. The indicated clearing price is the value that the clearing price would be if the book were to be frozen and the market called at that moment; it is the value that would maximize the number of shares that execute.

**indifference curve** In a two-good environment (X and Y), the locus of all combinations of X and Y that yield the same total utility. If constrained to a single indifference curve, the decision maker is indifferent about the specific point on the curve that he or she is at. Indifference curves for a utility mapping are analogous to the contour lines of a topographical map.

**inelastic price demand** A percentage change in price induces a less than proportionate percentage change in the quantity demanded.

**informational efficiency** The accuracy with which prices that are set in a marketplace reflect the underlying information that participants' expectations are based on.

**initial public offering** (IPO) The first public sale of equity shares by a previously privately owned enterprise.

**inside information** Information possessed by a select set of people who have a special access to information. Corporate officers and others with business ties to a company (such as an investment banker, lawyer, or proofreader) are commonly in possession of "inside information." Trading based on inside information is illegal.

**inside spread** The inside market. The difference between the lowest of all of the ask quotes, and the highest of all of the bid quotes.

**inter-dealer trading** Dealer-to-dealer trading. Generally motivated by a dealer seeking quickly to rebalance an inventory position.

**interdependent valuations** Investors' reassessments of their individual valuations that are based on what they know about others' valuations. See "adaptive valuations."

**isocost** A curve that shows how two different factors of production can be combined in different proportions at the same total cost. Each isocost curve is the locus of all of the different labor and capital combinations that require the same level of total expenditure.

**isoquant** A curve that shows how two different factors of production can be combined in different proportions that yield equal amounts of output. Each isoquant is the locus of all of the different labor and capital combinations that result in the same level of output. Isoquants in production theory play a similar role as indifference curves in consumption theory.

## K

**kinked demand curve model** A model of oligopolistic competition. The model implies that administered prices will be relatively constant in an oligopoly industry (as they are observed to be).

## L

**limit order** A priced order to buy or to sell a specific number of shares of a stock. For a buy order, the price limit is a maximum. For a sell order, the price limit is a minimum.

**liquidity** The characteristic of a micro market that permits buyers and sellers to trade reasonably quickly, in reasonable amounts, at reasonable prices. An attribute of the shares of an asset that, along with risk, affect the expected return and hence the price of an asset.

**long-run average costs** A firm's average cost of output when all inputs are freely variable. Total long run costs divided by output.

**long-run marginal costs** The change in a producer's long-run total cost with respect to a change in its level of output when all factors of production are freely variable (as they are in the long run).

## M

**macroeconomics** The economics of national or regional production and consumption emphasizing the formulation of economic policies. Macroeconomics stresses the importance of flow variables such as aggregate income, production, employment and interest rates, while microeconomics stresses the importance of variables such as relative prices.

**marginal rate of substitution** In a two-good model (*X* and *Y*), the slope of an indifference curve as given by the negative of the ratio of the marginal utility of *X* to the marginal utility of *Y*. The ratio of marginal utilities is the rate at which a consumer would be willing to trade one good for another, while neither augmenting nor diminishing his or her utility.

**marginal rate of technical substitution** the rate at which a producer can substitute one input for another while keeping output constant. In a two-input environment, the marginal productivity of one input divided by the marginal productivity of the other input. The negative of the slope of an isoquant.

**marginal utility** The change of utility obtained from consuming a good or service with respect to the change in the amount of the good or service consumed, all else equal.

**marginal values** The change in the value of a depend variable with respect to the change in the value of an independent variable. For instance, with the total cost of production being an increasing function of the total amount of a good or service produced, the marginal cost of production is the amount by which the total cost increases as output increases.

**market architecture** The design of a trading venue, including the rules and regulations that governing trading.

**market bid-ask spread** The best (lowest) posted market ask minus the best (highest) posted market bid.

**market breakdown** (failure) When a specific market fails to achieve an efficient market outcome from a public policy point of view. The failure can be in terms of price established and/or quantity traded. It can be attributable to factors such as externalities, asymmetric information, or moral hazard problems. In the extreme, a market can actually shut down (that is, totally ceases operations).

**market information** Information about current market conditions including current market quotes, transaction prices, and transaction volumes.

**market makers** Dealers with an affirmative obligation to continuously make two-sided markets (i.e., to continuously post bid and offer quotes).

**market order** An unpriced order to buy or to sell a specific number of shares of a stock at whatever the market price might be.

**market portfolio** In the frictionless market environment of CAPM, the single portfolio that contains all of the stocks that are traded in the market. Every individual in a CAPM environment achieves an optimal risk-return combination by combining shares of the market portfolio with the risk-free asset according to his/her individual tastes for risk. See also "separation theorem."

**microeconomic theory** The study of how markets operate. A common definition is the analysis of the allocation of scarce resources between competing ends.

**micro market** The market for a particular product (good or service).

**monopolistic competition** Characterizes markets where barriers to entry are low, and enough firms are in the industry so that no two firms will strategically respond to each others' pricing and output decisions. The outputs of the competing firms are somewhat differentiated under monopolistic competition (while they are fungible under perfect competition).

**monopoly** An industry where one firm is the sole producer/supplier of the product.

**monopsony** An industry where one firm is the sole buyer of the product.

**moral hazard** A situation that exists when a decision maker who does not bear the full responsibility for all possible outcomes acts more recklessly than he or she otherwise would and thereby imposes a cost on others.

**movements along a curve** For instance, the coordinated change in the price of a good and the quantity of the good that is purchased, all else held constant.

## N

**Nash equilibrium** In a gaming situation, an equilibrium that is reached by a small number of contestants who, because they do not cooperate with each other, fail to achieve the larger positive payouts (or lesser negative payouts) that would have been attainable if they had cooperated. A formulation that is applied (among other things) to oligopolistically competitive firms. The model was set forth by the Nobel laureate John Nash.

**negative externalities** An action involving the production, exchange, or consumption of a good or service that imposes a cost on others who are not directly involved in the action.

**network externality** When the users of a product can be viewed as having formed a network, and where the value of the network for the users increases with the number of members in the network. No one participant, in deciding whether or not to join a network, takes into account the value that his or her joining would have for others, which accounts for the use of the term "externality." Fax machines are a good example of a network (the more people who have a fax machine, the more people anyone with such a machine can send a fax to), but nobody while buying a fax machine takes account of the benefit that his/her purchase will bestow on others.

**nondiversifiable risk** That portion of the total returns variance for an asset that is correlated with the returns on the market portfolio and thus cannot be eliminated by portfolio diversification.

**normative economics** An economic formulation that is purposed to prescribe how an individual participant (household or firm) ought to respond.

## O

**oligopoly** An industry that comprises a few enough number of firms for the firms that are in the industry to respond to each other's competitive decisions (e.g., pricing, product innovation, and advertising decisions).

**optimal amount** Neither the maximum nor the minimum, an optimal amount is just the right amount. For instance, in a two-good universe ($X$ and $Y$), the optimum amount is the best quantity of $X$ to consume when more or less of good $X$ can be purchased. For a household, when the inputs of all goods ($X$, $Y$, etc.), have been *optimized*, the decision maker's utility has been *maximized*. Similarly, a firm's profits are maximized when its factors of production (labor and capital) are being used in optimal amounts.

**order-driven market** A market where public participants trade with each other without the intervention of a dealer (market maker). Limit orders posted by some public participants establish the prices at which other public participants can trade by market order.

**over-the-counter market (OTC)** A decentralized market for securities. Participants in an OTC market typically trade with a dealer firm that they communicate with either by telephone or computer. The small-cap, less liquid segment of the broader NASDAQ market.

**own price elasticity** The percentage change in the quantity demanded or supplied for a product divided by the percentage change in that product's own price, all else constant.

## P

**path dependent** The final, sustainable value that a variable (such as price) converges on depends on the specific path that it has followed during the earlier part of its price discovery process.

**pennying** The practice of placing a bid one penny above the highest existing bid on the book, or of placing an offer one penny below the lowest existing offer on the book so as to gain priority over the previously posted orders.

**perfect competition** A market that meets the conditions of (1) the product produced by the industry is, across all firms, identical (fungible), (2) many buyers and sellers are in the industry, and (3) no long-run barriers to entry or exit. Each firm in a perfectly competitive industry is "atomistic." Being atomistic, no individual firm in a perfectly competitive industry has any power to affect the price of the product that defines the industry.

**Ponzi scheme** A fraud scheme by which an investment company uses funds from new investors to pay preexisting investors. Also known as a pyramid scheme.

**positive economics** An economic formulation that is purposed to predict (as opposed to prescribe) how a representative market participant (or all participants in aggregate) will respond to a change in an economic variable.

**positive externalities** An action involving the production, exchange, or consumption of a good or service that benefits others who are not directly involved in the action. Some externalities can be internal to an industry but external to the individual firms in the industry; such economies lower the average-cost curves for the individual firms as the industry expands (for instance, a larger industry may, through a central organization, be able to provide better information at lower cost to the individual firms that are its members).

**preferencing** a common practice in equity trading that refers to an order being sent to (preferenced to) a specific dealer (or market) regardless of the price that is being quoted by other dealers (or markets). Preferencing occurs because of a special relationship or arrangement with that dealer (or market). The dealer firm to whom an order has been preferenced will typically match the best quote currently posted on the market.

**price ceiling** An externally imposed, maximum price. A ceiling is operative if placed below a free market equilibrium price.

**price discovery** In any micro market, price discovery is the process of finding an equilibrium value for a good, service, or productive resource that is being traded. Price discovery is a complex, imperfect process when participants do not simultaneously reveal their complete demand and supply functions. In an equity market, price discovery is the dynamic process of finding a value that best reflects the broad market's desire to hold shares of a stock. In equity markets, price discovery occurs most prominently after the receipt of news and following market openings as prices adjust to new economic conditions and to changes in the investors' desires to hold shares.

**price discrimination** A firm's practice of segmenting its market and selling the same product to different customers at different prices. Relatively high prices are charged in market segments where demand is relatively price inelastic, and relatively low prices are charged in market segments where demand is relatively price elastic.

**price floor** An externally imposed, minimum price. A floor is operative if placed above a free-market equilibrium price.

**price improvement** Providing a customer with a better price than the price that has been publicly quoted. Executing a market buy order at a price lower than the lowest quoted offer, and executing a market sell order at a price higher than the highest quoted bid.

**price priority** An order execution rule that specifies that the most aggressively priced order (the highest bid or the lowest ask) executes first.

**price taker** A market participant who, being of insignificant size in the market, has no power to affect the price that is set in the marketplace. Having no power over price, such a participant may be referred to as being "atomistic."

**primary markets** The market for the offering of new shares of a specific equity (common stock). In secondary market trading, the term is also used to designate the main market where the stock trades when there are several other, smaller markets that orders can be sent to.

**principal-agent problem** A problem that arises when, due to incomplete contracting and imperfect monitoring, an agent's own interests are in conflict with those of the principal who has hired him/her.

**prisoners' dilemma** A special case of a competitive (typically two-party) game that results in a Nash equilibrium. So called because it represents the strategic behavior of two individuals who, having been apprehended for a crime, are being interrogated by the police in separate rooms (so that the two prisoners cannot coordinate their stories).

**private information** Stock-related information that individuals possess because of their own investigations and analyses of the publicly available information set.

**producer surplus** The (larger) total amount that a producer (supplier) receives for selling *X* units of output less the (smaller) total amount that the producer (supplier) would have been willing to receive. Producer surplus is realized when the market price of *X* is greater than the producer's reservation price for supplying *X*.

**production function** The functional relationship between the inputs used by a firm and the output of its product.

**public goods** Goods that nonusers cannot be excluded from consuming. Goods for which one person's consumption does not reduce the availability of the good for others. A lighthouse in the harbor is a classic example of a public good; all ships passing can pick up the signal, and one ship picking up the signal does not lessen the signal that other ships can receive.

**public information** Widely disseminated information that all public participants can obtain.

## Q

**quantity discovery** The process of buyers' orders and sellers' orders meeting in the market and being turned into trades. The process is complex and incomplete when buyers and sellers do not reveal their full trading intensions because they are afraid that doing so will cause adverse price movements. When quantity discovery is incomplete, a latent demand to buy and to sell exists.

**quants** Market participants who use advanced mathematical techniques to value assets, formulate investment strategies, make trading decisions, and so on.

**quotation** A price at which someone is willing to buy or to sell shares, and the number of shares that the quote is good for. A quote can be either "firm" or "indicative."

**quote-driven market** A market where public participants do not trade directly with each other, but with dealers (market makers) who, by posting their quotes, establish the prices at which others can trade.

**quote matching** Upon receiving a preferenced order from a customer, a dealer firm will generally execute the customer's order at the current best quote in the market (highest bid or lowest offer) regardless of what the dealer firm may itself have been quoting at the time it received the order.

## R

**random walk** When the next step that an individual takes is totally independent of all of the steps that he or she has been previously taken, the person is said to be taking a "random walk." The term is used with regard to the dynamic behavior of stock prices when the next return for a stock (the stock's next "step") is not correlated with the stock's past returns (previous "steps"), and the next price change for a stock is totally unpredictable.

**rectangular hyperbola** A functional relationship between two variables, $X$ and $Y$, where their product, $XY$, is a constant (e.g., $Y = aX^{-1}$). Average fixed costs ($Y$) as a function of output ($X$) is an example of a rectangular hyperbola.

**Reg. FD** The Securities and Exchange Commission's 2000 rule that any listed company must immediately publicly disclose any information once the information has been provided to any public participant.

**regulations** The rules established by external, government regulators.

**regulatory arbitrage** A regulated firm exploiting competing regulators so as to obtain the most favorable regulatory oversight.

**regulatory lag** A delayed regulatory response to innovation, technological change, changing economic circumstances, or any other development affecting a micro market, or a delayed regulatory response to an industry's or a firm's request for a new regulatory rule or facility.

**representative consumer or supplier** An individual participant in the market who, in playing the role of an average, represents all participants. Predicting the response of a representative participant to the change in an economic variable is equivalent to predicting how all participants in aggregate will behave.

**reservation price** The maximum price a customer would be willing to pay for a given quantity when the alternative is to obtain nothing at all. It is also the minimum that a seller would be willing to receive for a given quantity when the alternative is to sell nothing at all.

**returns to scale** The rate at which a firm's output increases as inputs are increased proportionately. If equal movements out along a firm's expansion path result in increasing increments of output, the firm is operating with increasing returns to scale; if constant increments of output are realized for equal movements along the path, the firm is operating under conditions of constant returns to scale; and if decreasing increments of output are being realized, the firm is experiencing decreasing returns to scale.

**risk and return trade-off** When return has been maximized for a given level of risk and risk has been minimized for a given level of return, the amount of additional risk that has to be accepted in order to achieve a higher expected return or, alternatively viewed, the lower expected return that has to be accepted in order to achieve a lower level of risk.

**risk and uncertainty** In a risk situation, the distribution of possible outcomes and their associated probabilities are known. That is, with risk, the parameters of a distribution (such as its mean and variance) are known. Under uncertainty, the distribution of outcomes and their associated probabilities are not known. Uncertainty can in principal be reduced by information collection; risk cannot be reduced by gathering more information per se, but it can be controlled by appropriate hedging and portfolio formation.

**risk premium** The positive difference between the expected return for a risk position and a risk-free return that is required to induce an investor to accept taking the risky position.

**risk-free asset** a riskless asset such as a Treasury bill that with certainty yields a positive rate of interest. The risk-free asset plays an important role in the frictionless world of CAPM, where an individual investor can obtain an optimal risk, return configuration by combining the risk free asset with shares of the market portfolio according to his or her tastes for risk. See also "separation theorem."

**rules** Uniform and standard practices and methods typically established by an SRO, and formally approved by the overseeing government regulator.

## S

**scaled order** An order that is broken into two or more smaller pieces that are entered at two or more different prices. The submission of a downward sloping buy curve (or a subset of points along the buy curve) is a scaled order.

**secondary market** The market where already issued equity shares are traded (for instance, the New York Stock Exchange, NASDAQ, the London Stock Exchange, or Deutsche Börse).

**self-regulatory organization (SRO)** A nongovernment regulatory entity that establishes and enforces rules of conduct for a firm or industry group that it has responsibility for.

**sell-side intermediaries** Brokers and dealers who sell trading services to their buy-side customers.

**separation theorem** A theorem in financial economics that states that the optimal combination of stocks in a risky market portfolio (i.e., the structure of the market portfolio) is the same for all investors regardless of their individual risk, return preferences. The separation theorem holds in a CAPM environment.

**shift of a curve** Consider, for instance, the undefined consumption good, $X$. Let the demand for $X$ be a function of the price of $X$, the price of a related good (call it $Y$), income, and the consumer's tastes as described by a utility function. A change in the price of $X$ with all other relevant variables constant (price of $Y$, income, and tastes) results in a move along the consumer's demand curve for $X$. A change in any of the other variables (either the price of $Y$, income, and/or tastes) with the price of $X$ constant, results in a shift of (not a movement along) the consumer's demand curve for $X$.

**short selling** Selling borrowed stock on the expectation that a future decline in the stock's price will permit a subsequent (and profitable) purchase at a lower price.

**short-run average costs** A firm's average cost of output when, in the short run, one or more factor inputs is fixed. Total short-run cost divided by output.

**short-run marginal costs** The increase in a firm's total cost with respect to an increase in output when, in the short run, one or more factor inputs is fixed.

**socially desirable results** Economic outcomes that optimize the allocation of scarce resources.

**standard deviation** A measure of dispersion that is applied to normal distributions. The square root of variance.

**stochastic process** A process whereby a variable obtains values that are nondeterministic (i.e., have a random component.)

**stock variable** A variable that is measured as of a point in time (in contrast with a flow variable that is measured as a rate per unit of time). Wealth, an inventory position, and a stock price are examples of a stock dimensioned variable.

**strategic trading decisions** The order placement decisions made by traders seeking to optimize their gains from trading.

**substitutes** Products that are substituted for one another as their relative prices change. If *X* and *Y* are substitutes, the cross elasticity of demand for *X* with respect to the price of *Y* is positive. Tea and coffee, for instance, are substitutes. With perfect substitutes, only *X* or only *Y* is generally consumed, depending on which is less expensive.

**supplemental liquidity** Liquidity that is provided to a market by market makers over and above that which the market would freely supply to itself.

**supply curve** A graphic representation of the relationship between the quantity of a product supplied by a firm or an industry, and the price of the product.

**systematic risk** See "nondiversifiable risk."

**systemic risk** The risk of a widespread breakdown of markets and economic infrastructure.

## T

**technical analysis (charting)** The use of recent price data to predict future price movements.

**tick size** The minimum price variation; the units in which prices are stated (e.g., pennies).

**time priority** A rule of trading that specifies that, if two or more orders are tied at the most aggressive price, the order that has been placed first execute firsts.

**trading the order flow** Short-term trading by, e.g., a market maker or day trader, that is motivated not by fundamental information concerning a stock, but by the stock's dynamic, short-period price behavior. Buying because prices appear to be temporarily depressed, or selling because prices appear to be temporarily elevated, and then trading out of the position once prices have reverted to more appropriate levels.

**trading with immediacy** Submitting an order to buy or to sell and receiving an immediate execution. A submitted market order that will trade against a posted limit order or a market maker's quote will generally execute with immediacy.

**two-sided quotes** An agent such as a securities dealer simultaneously posts both a bid quote at which he or she will buy shares, and an ask quote at which he or she will sell shares.

## U

**unbalanced inventory** The accumulation of an unacceptably long position or an unacceptably short position in an asset (e.g., stock or bond).

**upstairs market** An off-exchange market (typically the "upstairs" offices of a bank or brokerage house) where trades can be made.

**utility value** The value to a consumer of a good or service measured in terms of the pure pleasure obtained (or pain avoided). Utility is an abstract, theoretical concept that is represented by ordinal, not cardinal numbers.

**V**

**variable costs** The costs of producing a good that vary with the amount of the good produced. In the long run, all costs are variable.

**W**

**wishful demand** The amount of a good that a buyer would ideally like to consume but cannot necessarily obtain, because he or she does not have the resources to make the purchase.

# About the Author

**Robert A. Schwartz** is Marvin M. Speiser Professor of Finance and University Distinguished Professor in the Zicklin School of Business, Baruch College, CUNY. Before joining the Baruch faculty in 1997, he was Professor of Finance and Economics and Yamaichi Faculty Fellow at New York University's Leonard N. Stern School of Business, where he had been a member of the faculty since 1965. In 1966, Professor Schwartz received his Ph.D. in Economics from Columbia University. His research is in the area of financial economics, with a primary focus on the structure of securities markets. He has published over 60 refereed journal articles, 12 edited books, and five authored books, including *The Equity Trader Course*, coauthored with Reto Francioni and Bruce Weber (John Wiley & Sons, 2006), *Equity Markets in Action: The Fundamentals of Liquidity, Market Structure, and Trading,* co-authored with Reto Francioni (John Wiley & Sons, 2004), and *Reshaping the Equity Markets: A Guide for the 1990s* (Harper Business, 1991; reissued by Business One Irwin, 1993). He has served as a consultant to various market centers, including the New York Stock Exchange, the American Stock Exchange, NASDAQ, the London Stock Exchange, Instinet, the Arizona Stock Exchange, Deutsche Börse, and the Bolsa Mexicana. From April 1983 to April 1988, he was an associate editor of the *Journal of Finance*, and he is currently an associate editor of the *Review of Quantitative Finance and Accounting,* and the *Review of Pacific Basin Financial Markets and Policies,* and is a member of the advisory boards of *International Finance* and *The Journal of Trading.* In December 1995, Professor Schwartz was named the first chairman of NASDAQ's Economic Advisory Board, and he served on the EAB until spring 1999. He is developer, with Bruce Weber and Gregory Sipress, of the trading and market structure simulation, TraderEx (www.etraderex.com/). In 2009, Schwartz was named the first recipient of the World Federation of Exchanges' annual Award for Excellence.

# Index

Active traders, 139–140
Adaptive valuations, 311, 327, 337–338
Adjustment costs, 136–142
Adjustment path, 64
Administered prices, 148, 253, 264–268, 285
Aggregation:
  description of, 81
  from individual demands to market demand, 97–98
  from individual trade curves to market curves, 113–116
Aggressive quotes, competition by, 278–280
Algorithmic trading:
  conceptual foundation for, 326–327
  description of, 324–325, 367
  intraday volatility and, 330
Allocation efficiency, 293, 295–298
Alternative Trading Systems (ATS), 392
American International Group (AIG), 333, 338, 341
Analysis:
  based on fundamental information, 308
  dynamic and static, 64
  long-run and short-run, 14–15
  marginal, 15–18
  of supply propensities of industry, applicability of, 230–231
  *See also* Demand-supply analysis of markets; Technical analysis
Antitrust regulation, 399–400
Arbitrage, regulatory, 358
Arbitrage trading, 139, 331
Arc elasticity, 130
Archipelago ECN, 356, 392
Architecture, definition of, 5
Ask quotations, 138
Assumptions:
  of CAPM, 121–122
  theory and, 5
Asymmetric information and market failure, 299–300, 337
ATS (Alternative Trading Systems), 392
Auto manufacturers, 341
Automation, impact of, 363–368
Average cost curves, 197–200, 201–205
Average fixed costs, 200–201
Average risk premium, 52

Back-office crisis, 371–372
Bailout programs, 339
Bandwagon effect, 124–126
Barriers to entry to monopoly power, 257
Baruch, Bernard, 332–333
BBO (best bid and offer), 138
Behavioral economics, 329–330
Bernstein, Peter, 143
Best execution criteria, 383–386
Best profit output, determining, 206–208
Bid-ask spreads:
  in competitive dealer markets, 277–278
  description of, 138, 140
  in order-driven markets, 154–155
  as price of dealer services, 272
  in securities markets, 180–183
Bid quotations, 138
Bids, 154
Block markets, 164
Book building, dynamics of, 271
Books, 154. *See also* Limit-order books
Brand, and competition, 282

Budget constraint:
  capital market line compared to, 56–57
  in consumer choice model, 29–31, 39
  isocost line compared to, 190
Buttonwood Tree Agreement, 386
Buy curves, 111–116, 174–175
Buy-side players, 149

Call auctions:
  description of, 127
  large traders' dilemma and, 270–271
  in order-driven markets, 160–163
  strategic order placement in, 174–176
Capital Asset Pricing Model (CAPM):
  demand and supply analysis and, 117–122
  description of, 58–59
  homogenous expectations and, 310
  market efficiency and, 312
  perfectly liquid environment and, 143
Capital market line, 54–58
Cap size, 233–234
Cardinal numbers, 31
*Ceteris paribus* concept, 11–13
Charting/chartists, 316, 324–325
Clearing price, 162–163
Collateralized debt obligations, 334
Commissions, fixed, 378–383
Commodities Futures Trading Commission, 373
Comparative statics:
  current consumption, future consumption trade-off and, 67–70
  description of, 15, 64
  risk, return trade-off and, 70–73
  *X, Y* trade-off and, 64–67
Competition:
  duopolistic, 268–270
  in equity markets, 270–281
  free markets and, 7, 9
  government intervention and, 399–401
  imperfect, 242, 260–270, 285
  large traders' dilemma, 270–271
  monopolistic, 260–263, 399–400
  NYSE and, 282
  oligopolistic, 263–270
  overview of, 241
  price and, 281–282
  in regulation, 358
  *See also* Monopoly firm; Perfect competition
Competitive dealer markets, 277–281, 347–350
Competitive structure of dealer markets, 224–225
Complements, 90–96
Computerized trading, 304–305
Constant returns to scale, 187
Consumer choice model:
  budget constraint, 29–31
  current consumption, future consumption trade-off, 35–43
  description of, 28–29
  indifference curves, 31–33
  optimal resolution of *X, Y* trade-off, 33–34
  risk, return trade-off and, 43–63
  *X, Y* trade-off, 28–34
Consumer surplus, 166–168, 248
Consumption, as flow variable, 36–37. *See also* Current consumption, future consumption trade-off
Contestability/contestable markets, 257
Continuous-order-driven markets, 153–160, 173–174
Contract curves, 297
Convexity of indifference curves, 60
Coordination of orders, 145–146
Corporate equities, 3
Cost control, and dealer markets, 225–226
Cost curves:
  average cost, 197–200, 201–205
  average fixed costs, 200–201
  dealer markets, 226–228
  for firms in competitive industry, 243
  long-run and short-run, 187, 201–205
  long-run total, 193–195

short-run total, 195–197
*See also* Marginal cost curves
Costs:
execution, 137, 140–142
explicit, 137
fixed and variable, 13–14
of ignorance, 218–219, 222
market maker, 218–219
of trading, 137–142
of unbalanced inventory, 218, 221–222
*See also* Cost curves
Covariance, 118–119, 120–121
Creative destruction, 336–337
Credit default swaps, 334, 339
Crossing orders, 154
Cross-price elasticity of demand, 86, 90–96
Current consumption, future consumption trade-off:
budget constraint, 39
comparative statics and, 67–70
description of, 35
income and consumption patterns, 36–37
indifference curves, 39–40, 42–43
interpretation, 40–43
price of current versus future consumption, 37–39
Customers, segmentation of, and price discrimination, 253–256

Deadweight loss, 247–249, 263, 277
Dealer markets:
cost curves, 226–228
description of, 215–216
dynamic inventory control, 223–224
dynamic pricing, 221–223
externalities and supply of services, 228–230
market maker operations, 216–219
oligopoly and, 272
output and cost control, 225–226
pricing models, 219–221
structure of, 149–153, 224–225, 277–281
Decisions of representative consumers, 82–84
Decreasing returns to scale, 210
Demand:
cross-price elasticity of, 90–96
effective compared to wishful, 11
elasticity of, 117–122
to hold shares of risky asset, 99–100
income elasticity of, 86, 87–88
intersection with supply, 104–105
own price elasticity of, 88–90
Demand curves:
aggregate market, 141–142
description of, 81
from individual to market, 96–99
investor's, for risky asset, 100–102
kinked model, 264–267
for monopolistically competitive firm, 262–263
of monopoly firm, 250–251, 259
ordinary and reservation, 165–166
of price-discriminating monopolist, 255
short-run versus long-run, 134–135
as tool of economic analysis, 12–13
for *X,* 84–86
Demand functions, 85
Demand-supply analysis of markets:
for risky assets, 109–112
for *X,* 102–108
Demutualization of exchanges, 355–357
Deutsche Börse, 4, 356
Diseconomies of scale, 187, 194–195
Dissemination of information, dynamic efficiency of, 320–322
Divergent expectations, 57, 147–148, 310–311
Divergent valuations, 327
Diversifiable risk, 118, 218
Dividend discount model, 326–327
Dodd, David, 331
Dollar size of spreads, 218
Dollar strength of opinion, 319

Duopoly, 263–264, 268–270
Dynamic analysis, 64
Dynamic efficiency of information dissemination, 320–322
Dynamic environment, and monopoly power, 256
Dynamic equilibrium, 15
Dynamic inefficiency, 321
Dynamic inventory control, 223–224
Dynamic pricing policy, 221–223
Dynamic programming model, 222

ECN (Electronic Communications Network), 392
Ecologies, equity markets as, 11, 400–401
Economic rents, 245–247
Economies of scale:
  average costs and, 200
  dealer services and, 228–230
  description of, 187
  exploitation of, 194–195
  monopoly firm and, 256–260
  as technological property of production function, 259
Economies of scope, 258
Edgeworth box, 296
Efficiency of markets, *see* Informational efficiency; Market efficiency; Price discovery efficiency
Efficient Market Hypothesis (EMH), 322–325, 331, 332
Elasticity:
  arc, 130
  of demand for stocks in portfolio, 117–122
  description of, 18–19, 81
  of income, 87–88
  market demand curve and, 98–99
  as measure of responsiveness, 86–87, 99
  monopolist and, 252
  own price, 88–90
  supply, 103
Electronic Communications Network (ECN), 392
Electronic trading platforms, 356
EMH (Efficient Market Hypothesis), 322–325, 331, 332
Envelope curves, 204, 205
Equality on margin, 231
Equilibrium:
  definition of, 15
  information-gathering/-production and, 317–318
  of market, 112–116
  in market for shares of risky asset, 109–112
  monopolist and, 252
  under monopolistic competition, 261–262
  in monopoly dealer market, 273–277
  Nash, applications of, 268–270, 271
  oligopoly and, 264–268
  perfect competition and, 242–244
  unique, 297–298, 328
Equilibrium in market for *X:*
  intersection of demand and supply, 104–105
  regulatory intervention and, 105–108
  supply curve of *X,* 102–103
Equilibrium prices, discovery of in competitive markets, 347–350
Equity markets:
  best execution criteria and, 383–386
  competition in, 270–281
  description of, 3, 132–133
  as ecologies, 10–11, 400–401
  electronic technology in, 365–368
  fixed commission structure and, 378–383
  implicit collusion and, 391
  informational efficiency and, 306
  largest of, 4
  large traders' dilemma, 270–271
  as networks, 258, 328
  as nonfrictionless, 145
  off-board trading restrictions and, 386–391

  penny prices in, 280–281, 393–395
  quality issues in, 397–401
  regulation of, 375–397
  trade-through rule and, 395–397
  trading costs, 137–142
  trading in as adjustment process, 137
  *See also* Dealer markets
European Commission, Markets in Financial Instruments Directive, 390–391
Exaggerator category of investors, 315
Excess returns, 245, 246
Exchange-based trading, 365
Exclusivity, and monopoly power, 257
Execution costs, 137, 140–142
Existing information set, and informational efficiency, 312–313, 314
Expansion path, 191–193
Expectations:
  divergent, 57, 147–148, 310–311
  homogenous, 57, 121, 310
  information and, 305–310
Expected average cost, 226–227
Expected marginal cost, 226–227
Expected number of transactions, 227–228
Expected return, risk, and return trade-off, 53–54
Expected utility, 28–29, 62, 176
Explicit costs, 137
Ex post regret, 219
Externalities:
  market failure and, 301, 338
  supply curves of firms and, 211–212, 228–230
  *See also* Network externalities
Extramarginal firms, 245

Factors of production, 3, 189, 293–294
Fails to deliver, 372
Fair bets, 46
"Fairness" argument, 400
Federal Deposit Insurance Corporation, 377
Financial capital, 294, 335
Financial Services Authority (FSA, U.K.), 384–386, 409
Financial turmoil of 2007-2009, 9, 333–341
Fish, optimal schooling behavior of, 234–237
Fixed commission structure, 378–383
Fixed costs, 13–14
Flat inventory position, 152
Flow-dimensioned demand, 137
Flow variables, 29, 36–37
Free market stance, 6–10, 105
Frictionless markets:
  description of, 5–6, 20–21
  information and, 306
  monopoly dealer model, 276
  *See also* Nonfrictionless markets
Frictions problem, generality of, 341–343
FSA (Financial Services Authority, U.K.), 384–386, 409
Fundamental information, 307–310

Gains from trading, 164–165
Game theory, noncooperative, 268–269
Gathering information, and informational efficiency, 313–318
Glass-Steagall Act, *see* Securities Act of 1933
Goldman Sachs, 333–334, 338
Good or service *(X)*:
  market efficiency for, 322–325
  ordinary and reservation demand curves for, 165–166
  production function for, 188–193
  short-run demand and supply for, 133–135
  *See also* Equilibrium in market for *X*
Governance structure of exchanges, 355–357
Government bailout programs, 339

Government regulation:
  caveats for, 397–401
  of financial markets, 355
  problems with, 373–375
  self-regulation and, 359–360
Graham, Benjamin, 331
Great Depression, 375–376

Herding, 125
High frequency data, 331–332
Homogenous expectations, 57, 121, 310
Honesty as regulatory goal for securities markets, 370–371

Ignorance, and financial turmoil of 2007-2009, 334–335
Immediacy:
  common price charged for, 224–225
  market maker firms and, 272
  as product, 217
Imperfect competition:
  description of, 242, 285
  monopoly, 260–263
  oligopoly, 263–270
Imperfect information:
  government regulation and, 375
  market failure and, 299–300, 337
Implicit collusion, 391
Income, as flow variable, 36–37
Income effect, 66, 67
Income elasticity of demand, 86, 87–88
Income expansion path, 87, 88
Indicated clearing price, 162–163
Indifference curves and allocation efficiency, 295–297
Indifference curves in consumer choice model:
  current consumption, future consumption trade-off, 39–40, 42–43
  risk, return trade-off, 60–62
  *X, Y* trade-off, 31–33
Industry, definition of, 249–250
Inelastic responses, 142
Inertia in dynamic environments, and market failure, 303–305, 340
Inferring informational change from prices, 314–316
Inflection points, 199–200
Information:
  as factor of production, 293–294
  imperfect and asymmetric, 299–300, 337, 375
  monopoly on, 376–377
  stock prices and, 330–333
  types of, 307–310
Informational accuracy, 318–320, 322
Informational efficiency:
  description of, 305–307, 322
  dimensions of, 293, 311–322
  existing information set and, 312–313, 314
  information-gathering/-producing activities and, 313–318
  test of, 322–325
Information economics, 306
Initial public offering (IPO), 3–4
Inside information, 308–309, 321, 370, 376–377
Inside spreads, 138, 278, 280
Interdealer trading, 229, 272
Interdependent valuations, 311
Intermediaries, 149, 356
Interrelated utility functions, 76
Intraday volatility, 330
Intramarginal firms, 245
Inventories, of dealer firms, 275–276
Inventory bounds, of dealer firms, 225
Inventory control mechanisms, 152, 223–224
Investing compared to trading, 10
Investors:
  behavior of, 326–327
  categories of, 315
  efficiency and aggressiveness of, 320–321
  value of information bits to, 316–317

IPO (initial public offering), 3–4
Irrational behavior and market failure, 300, 337–338
Isocost lines, 190–193, 232
Isoquants, 189–193, 232

Jobber's turn, 217

Kinked demand curve model, 264–267

Labor contracts, 341–342
Legal frameworks for exchanges, 359
Lehman Brothers, 333, 335, 337, 340–341
Limit-order books:
  for call auctions, 160–161
  for continuous markets, 155
  for order-driven markets, 154, 158–159
Limit orders, 24–25, 138, 155–156
Liquidity:
  customers motivated by, 273
  definition of, 4, 131, 141–142
  financial capital and, 294
  financial turmoil of 2008 and, 339
  market structure and, 146–147
  order coordination and, 145–146
  perfectly liquid environments, 143–144
  in real environments, 144–145
  supplemental, 217
Logarithmic returns, 49
London Stock Exchange, 4, 356
Long positions, 138
Long-run analysis, 14–15
Long-run cost curves, 187, 201–205, 211–212
Long-run supply curves, 134–135, 209
Long-run total cost curves, 193–195

Macro markets, definition of, 1
Madoff, Bernard L., 371
Malkiel, Burton, 324, 331
Maloney Act Amendments, 361, 377
Marginal analysis, 15–18
Marginal cost curves:
  description of, 197–200
  in long-run and short-run, 201–205
  for monopoly firm, 259
  for oligopoly, 267–268
  as supply curves, 213–215
Marginal firms, 245
Marginal rate of substitution, 33
Marginal rate of technical substitution (MRTS), 190–191, 192
Marginal revenue curves:
  of monopoly firm, 250–251
  no-response and perfect response, 265, 267
  of price-discriminating monopolist, 255
Marginal risk premium (MRP), 53, 62–63
Marketable limit orders, 156
Market architects, 1–2
Market bid-ask spreads, 138, 140
Market demand curves, 97–99
Market efficiency:
  allocation efficiency, 295–298
  for broker/dealer services, 371–372
  CAPM and, 312
  financial turmoil of 2007-2009 and, 333–341
  information and, 330–333
  overview of, 293–294
  for price discovery, 373
  test of, 322–325
  *See also* Informational efficiency; Price discovery efficiency
Market failure:
  conditions of, 299, 337–341
  externalities and, 301, 338
  imperfect and asymmetric information and, 299–300, 337
  irrational behavior and, 300, 337–338
  moral hazard and, 302, 339
  network externalities and, 301–302, 338–339
  as partial or complete, 298–299

Market failure (*continued*)
  principal-agent problems and, 302–303, 339
  public goods and, 303, 339–340
  short-run myopia and, 300–301, 338
  systemic risk and, 305, 340–341
  technological inertia and, 303–305, 340
Market impact, 140–141
Market information, 307
Market maker firms:
  costs of, 218–219
  dealer firms as, 278–280
  operations of, 272–273
  revenues of, 217–218
  services of, 216–217
Market makers, definition of, 149
Market microstructure, 4–5, 133
Market orders, 138, 155–156
Market risk premiums, 51–52
Markets:
  complexity of, 1–2
  as ecologies, 10–11, 400–401
  as fragile, 335
  functioning of, 6–10
  macro, 1
  over-the-counter, 361
  primary and secondary, 3–5
  reengineering of, 369
  segmentation of, and price discrimination, 253–256
  types of, 1
  *See also* Equity markets; Frictionless markets; Market structure; Micro markets; Nonfrictionless markets
Market structure:
  continuous-order-driven markets, 153–160
  description of, 148–149
  government involvement with, 398–399
  liquidity and, 146–147
  negotiated (block) markets, 164
  quote-driven (dealer) markets, 149–153
Maximization of expected utility, 28–29
Mean, 49
Microeconomic theory:
  *ceteris paribus* concept, 11–13
  competition and, 9–10
  elasticity, 18–19
  equilibrium, 15
  fixed and variable costs, 13–14
  free market stance and, 6–10
  long-run and short-run analysis, 14–15
  marginal analysis, 15–18
  market microstructure and, 4–5
  maximum, minimum, and optimum, 19–20
  reality and, 2–3, 20
Micro markets:
  definition of, 1
  driving forces in, 27
  forces shaping, 353
  regulation of, 354
  *See also* Equity markets
Models:
  dividend discount, 326–327
  monopoly dealer, 273–277
  normative and positive, 82
  pricing, in dealer markets, 219–221
  *See also* Capital Asset Pricing Model; Consumer choice model
Momentum effect, 125
Monopolistic competition, 260–263, 399–400
Monopoly dealer model, 273–277
Monopoly firm:
  barriers to entry and, 257
  best profit price and output, 250–253
  definition of, 7
  industry, definition of, 249–250
  natural monopolies, 257–260, 284
  price discrimination and, 253–256
  problems with, 256–257
Monopoly on information, 376–377
Monopsonist/monopsony, 274, 275
Moral hazard and market failure, 302, 339

Mosaic Theory, 309
MPR (marginal risk premium), 53, 62–63
MRTS (marginal rate of technical substitution), 190–191, 192

Naked shorts, 138
Nasdaq Stock Market, 4, 356, 361–363
Nash equilibrium, 268–270, 271
National Market System (NMS), 395–396
Natural monopolies, 257–260, 284
Negative externalities, 301
Negotiated (block) markets, 164
Network externalities, 258, 301–302, 338–339
Networks, properties of, 328
New York Stock Exchange (NYSE Euronext):
  branding of, 282, 357
  description of, 4
  fixed commission structure and, 379
  Hybrid Market system, 365
  Nasdaq and, 362
  off-board trading rule of, 386–391
  order-consolidation rule of, 368–369
  privatization of, 356
19c.3 stocks, 387–388
NMS (National Market System), 395–396
Noncooperative game theory, 268–269
Nondiversifiable risk, 218
Nonfrictionless markets:
  competition and, 282
  description of, 5–6, 131, 176–178
  monopoly dealer model, 276–277
  price discovery efficiency and, 330
  as reality, 147–148
  *See also* Equity markets; Frictionless markets
No response demand curves, 265, 266
No response marginal revenue curves, 265, 267
Normative models, 82
Number of shares outstanding, 116

Off-board trading restrictions, 386–391
Offers, 154, 216
Oligopoly:
  administered prices, multifirm equilibrium, and, 264–268
  characteristics of, 263–264
  dealer markets and, 272, 277–281
  duopoly, 268–270
"One quantity" assumption, 170–172
On the margin:
  current versus future consumption and, 40–41
  making adjustments, 15–18
Opinion, dollar strength of, 319
Opportunity costs, 140–141
Optivar trading company, 367–368
Order-driven markets, 153–160, 173–174
Order flow, trading, 218, 330
Order Handling Rules, 392
Order placement strategy, 170–176
Orders:
  attributes of, 146
  coordination of, 145–146
  crossing, 154
  limit, 24–25, 138, 155–156
  market, 138, 155–156
  scaled, 171
Ordinal numbers, 32
Ordinary buy curves, 174–175
Ordinary demand curves, 165–166
Output, and monopoly firms, 250–253, 256
Overshooting, 338
Over-the-counter (OTC) markets, 361
Own price elasticity of demand, 86, 88–90

Passive traders, 139–140
Patents, 257
Path dependency, 328–329
Penny prices, trading in, 280–281, 393–395
Percentage spreads, 217

Perfect competition:
  deadweight loss and, 247–249
  description of, 242
  models of, 7
  no-profit, long-run, equilibrium solution, 242–244
  profits, economic rents, and, 244–247
Perfect complements, 93–95
Perfect response demand curves, 265, 266
Perfect response marginal revenue curves, 265, 267
Perfect substitutes, 93–95, 120
Periodic call auctions, strategic order placement in, 174–176
Philanthropy, 76
Ponzi schemes, 370–371
Positive externalities, 301
Positive models, 82
Preferencing, 273, 278, 279–280
Price:
  administered, 148
  as competitive variable, 281–282
  of current versus future consumption, 37–39
  of dealer services, floor under, 280–281
  importance of, 241
  as important variable, 12
  indicated clearing price, 162–163
  inferring informational change from, 314–316
  informational accuracy of, 318–320, 322
  interference in free setting of, 132
  monopoly firm and, 250–253, 256
  as rationing variable, 116
  reservation price, 169–170, 176
  responses to change in, 136–137
  socially desired, 260
  *See also* Price discovery; Price discovery efficiency
Price controls, imposed, 106, 260
Price discovery:
  in competitive markets, 347–350
  description of, 132
  divergent expectations and, 311
  liquidity and, 144–145, 146
  quality of, as regulatory focus, 373
  in quote-driven markets, 150–152
Price discovery efficiency:
  behavioral economics and, 329–330
  description of, 322, 325–326
  investor behavior and, 326–327
  three punchlines to story, 328–329
Price discrimination, and monopoly firms, 253–256
Price elasticity, *see* Liquidity
Price improvement, 272–273
Price priority, 279
Price setters, 7
Price takers, 7
Price theory, 86
Price variation, 154
Price volatility, 50
Pricing models in dealer markets, 219–221
Primary markets, 3–5
Principal-agent problems and market failure, 302–303, 339
Prisoners' Dilemma, 268–269
Private goods, 303
Private information, 309–310
Privatization of exchanges, 355–357
Producer surplus, 170, 248–249
Producing information, and informational efficiency, 313–318
Production function:
  economies of scale and, 259
  for *X,* 188–193
Profit output:
  determining best, 206–208
  for monopoly firm, 260
  for oligopoly, 267–268
Profits:
  economic function of, 247
  economic rents and, 244–247
  perfect competition and, 242–244

Propositioned limit orders, 138
Public goods and market failure, 303, 339–340
Public information, 308
Public policy:
  caveats for, 397–401
  monopoly firms and, 260
  regulation of micro markets and, 354
  tick size standardization and, 281, 393

Quantity discovery, 132, 145, 146
Quotations, 138
Quote-driven (dealer) markets:
  cost curves, 226–228
  description of, 215–216
  dynamic inventory control, 223–224
  dynamic pricing, 221–223
  externalities and supply of services, 228–230
  market maker operations, 216–219
  oligopoly and, 272
  output and cost control, 225–226
  pricing models, 219–221
  structure of, 149–153, 224–225
Quoted spreads, and realized spreads, 226
Quote matching, 273

Random equilibria environments, 328
Random walks, 323–325
Rational decision making, 310
Rationing variables, 105–106, 116
Raw information, 307
Realized spreads, and quoted spreads, 226
Rectangular hyperbola, 110
Reengineering of markets, 369
Regulation:
  approaches to, 409
  equilibrium in market for *X* and, 105–108
  evolving focus of, 368–369
  financial turmoil of 2008 and, 342–343
  price controls, imposed, 106, 260
  self-regulation, 353–354, 357–360
  unintended consequences of, 8–9
  of U.S. equity markets, 375–397
  *See also* Government regulation
Regulation FD (SEC), 321, 370
Regulatory arbitrage, 358
Regulatory capture, 360
Regulatory lag, 360
Relative prices, 15
Rents, 245–247
Representative consumers, decisions of, 82–84
Reservation buy curves, 174–175
Reservation demand curves, 165–166
Reservation prices, 169–170, 176
Resistance bounds, 329
Responsiveness, measures of, 85, 86
Returns:
  abnormal, 313
  excess, 245, 246
  fixed and variable, 13–14
  measurement of, 48–50
  risk and, 19–20
Returns to scale, 194–195, 210
Revenues, market maker, 217–218
Reversal behavior, 125
Risk:
  diversifiable, 118
  measurement of, 50–51
  nondiversifiable and diversifiable, 218
  return and, 19–20
  systematic, 117
  uncertainty compared to, 43–44
  unsystematic, 117, 118
  volatility and, 335–336
Risk, return trade-off:
  comparative statics and, 70–73
  description of, 43–44, 53–60
  indifference curves, 60–62
  interpretation, 62–63
  measurement of returns, 48–50
  measurement of risk, 50–51
  optimal solution to problem of, 62
  risk aversion, 44–48
  risk premiums, 51–53

Risk arbitrage trading, 139
Risk aversion, 44–48, 51
Risk-free assets, 117–122
Riskless arbitrage trading, 139
Risk premiums, 47, 48, 51–53
Risky assets, 99–102, 109–112
Round trips, 217
Rugged individualist category of investors, 315

Sales taxes, 107–108
Saving spreads, 156
Scaled orders, 171
Scale economies, *see* Economies of scale
Schering-Plough, 309–310
Schumpeter, Joseph A., 336
SEC, *see* U.S. Securities and Exchange Commission
Secondary markets, 3–5
Securities Act of 1933, 370, 376, 377
Securities Acts Amendments of 1975, 378, 383, 386, 388
Securities Exchange Act of 1934, 370, 376, 378
Securities exchanges:
  governance structure of, 355–357
  regulatory goals for, 370–373
  self-regulation of, 357–360
Security analysis based on fundamental information, 308
Segmentation of markets, and price discrimination, 253–256
Self-regulation of exchanges, 353–354, 357–360
Self-Regulatory Organization (SRO), 357
Sell curves, 111–116
Selling short, 138
Sell-side intermediaries, 149
Semi-strong form efficiency, 323
Separation theorem, 121
Service, *see* Good or service
Sheep category of investors, 315
Shelley, Percy Busse, 401
Short positions, 138
Short-run analysis, 14–15
Short-run cost curves, 187, 201–205
Short-run demand and supply for *X*, 133–135
Short-run myopia and market failure, 300–301, 338
Short-run total cost curves, 195–197
Short selling, 138
Slope of isoquant, 190
Smith, Adam, on invisible hand, 6
Snob effect, 124–126
SRO (Self-Regulatory Organization), 357
Stable value, 105
Standard deviation, 51
Static analysis, 64
Static efficiency, and dynamic efficiency, 321–322
Static equilibrium, 15
Stigler, George, 306
Stocks, elasticity of demand for, 117–122
Stock variables, 29
Strategic order placement, 170–176
Strategic trading decisions:
  consumer surplus, 166–168
  demand curves for, 165–166
  description of, 164–165
  reservation price, 169–170
  trader's objective, 168
Strong form efficiency, 323
Subsidies, 107–108
Substitutability, 96, 119–122, 134–135
Substitutes, 90–96, 249–250
Substitution effect, 66–67
Summers, Lawrence, 332
Supplemental liquidity, 217
Supply, intersection with demand, 104–105
Supply curves:
  for competitive firms, 208–215
  of competitive firms, 188
  of competitive industries, 211–212, 243
  marginal cost curve as, 213–215
  short-run and long-run, 134–135, 209
  of *X*, 102–103

Supply elasticity, 103
Support bounds, 329
Surplus, 170
Systematic risk:
  demand and supply analysis and, 117, 121
  liquidity and, 143–144
Systemic risk and market failure, 305, 340–341

Taste, and utility function, 133
Tax, on economic rents, 246
Technical analysis:
  behavioral economics and, 329–330
  conceptual foundation for, 326–327
  description of, 324–325
Technological inertia and market failure, 303–305, 340
Technology change:
  impact of automation, 363–368
  Nasdaq Stock Market and, 361–363
Theory:
  assumptions and, 5
  reality and, 2–3, 20
Tick size, imposition of minimum, 280–281, 393
Time priority:
  in electronic order-driven markets, 279
  pennying and, 394
  as rationing device, 163
"Too big to fail" concept, 338
Total risk premium, 52
Trade credit, extension of, 287–289
Trade curves, 111–116
*TraderEx: The Stock Market Trading Game,* 24
Trade-through rule, 395–397
Trading:
  algorithmic, 324–325, 326–327, 330, 367
  arbitrage, 139, 331
  automation of, 365–368
  computerized, 304–305
  exchange-based, 365
  gains from, 164–165
  interdealer, 229, 272
  investing compared to, 10
  order flow, 218, 330
  in penny prices, 280–281, 393–395
  *See also* Strategic trading decisions
Trading costs, 137–142
Two-sided quotes, 149

Uncertainty, 43–44, 336–337
Unintended consequences, 8–9
Unique equilibrium, 297–298, 328
Unsystematic risk, 117
U.S. Securities and Exchange Commission (SEC):
  honesty in markets and, 370
  Order Handling Rules, 392
  Regulation FD, 321, 370
  Regulation National Market System, 395–396
  Regulation of Exchanges and Alternative Trading Systems, 392
  reputation of, 377
  rules-based regulatory framework of, 409
  Securities Acts Amendments of 1975 and, 388–389
Utility function, 31, 47–48, 76, 133
Utility value, 16–17

Valuations of shares:
  adaptive, 311, 327, 337–338
  divergent, 327
  "fuzzy," 327
  interdependent, 311
Variable costs, 13–14
Variables:
  flow and stock, 29, 36–37
  rationing, 105–106, 116
Variance, 49, 50, 118–119, 120–121
Volatility:
  financial turmoil of, 340, 342
  intraday, 330
  risk and, 335–336

Weak form efficiency, 323
Wealth of decision makers, 319

*X* (good or service):
  market efficiency for, 322–325
  ordinary and reservation demand curves for, 165–166
  production function for, 188–193
  short-run demand and supply for, 133–135
  *See also* Equilibrium in market for *X*

*X, Y* trade-off:
  comparative statics and, 64–67
  consumer choice model and, 28–34
  cross-price elasticity and, 90–96
  demand curve for *X,* 84–86
  elasticity and, 86–87
  income elasticity and, 87–88
  own price elasticity and, 88–90

Praise for

# Micro Markets

"Robert Schwartz's *Micro Markets* is a lucid intermediate microeconomics textbook with a good sense of humor that makes it a pleasure to read. It is an excellent presentation of price theory, with numerous applications to financial markets, for students in economics and finance and for MBA students."

**—Barry R. Chiswick, Distinguished Professor, Department of Economics, and Director, Center for Economic Education, University of Illinois at Chicago**

"*Micro Markets* is a wonderful microeconomics text. Professor Schwartz's passion and talent for teaching shines through on every page."

**—Dr. Benn Steil, Senior Fellow and Director of International Economics, Council on Foreign Relations**

"*Micro Markets* is an excellent, fascinating book. With skill and penetrating insight, Professor Schwartz has united traditional microeconomic theory with the workings of real-world markets. In so doing, he has underscored the important effects that factors such as illiquidity, trading frictions, and imperfect knowledge can have on market outcomes. His book will have much appeal to a broad range of readers, including academicians, economics and finance students, and industry professionals."

**—Reto Francioni, CEO, Deutsche Börse Group, Frankfurt, Germany**

wileyfinance.com

ISBN 978-0-470-44765-9
59500
9 780470 447659